

09/19/00
JC803 U.S. PTO

Please type a plus sign (+) inside this box → ☐

PTO/SB/05 (4/98)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 C.F.R. § 1.53(b))

Attorney Docket No. 520.33330CC3

First Inventor or Application Identifier Masashi TSUCHIDA

Title See 1 in Addendum

Express Mail Label No.

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. ☒ * Fee Transmittal Form (e.g., PTO/SB/17)
(Submit an original and a duplicate for fee processing)
2. ☒ Specification [Total Pages 83] 1
(preferred arrangement set forth below)
- Descriptive title of the Invention
 - Cross References to Related Applications
 - Statement Regarding Fed sponsored R & D
 - Reference to Microfiche Appendix
 - Background of the Invention
 - Brief Summary of the Invention
 - Brief Description of the Drawings (if filed)
 - Detailed Description
 - Claim(s)
 - Abstract of the Disclosure
3. ☒ Drawing(s) (35 U.S.C. 113) [Total Sheets 23] 1
4. Oath or Declaration [Total Pages 2] 1
- a. ☐ Newly executed (original or copy)
- b. ☒ Copy from a prior application (37 C.F.R. § 1.63(d))
(for continuation/divisional with Box 16 completed)
- i. ☐ DELETION OF INVENTOR(S)
Signed statement attached deleting
inventor(s) named in the prior application,
see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

5. ☐ Microfiche Computer Program (Appendix)
6. Nucleotide and/or Amino Acid Sequence Submission
(if applicable, all necessary)
- a. ☐ Computer Readable Copy
- b. ☐ Paper Copy (identical to computer copy)
- c. ☐ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

7. ☐ Assignment Papers (cover sheet & document(s))
8. ☐ 37 C.F.R. § 3.73(b) Statement ☐ Power of
(when there is an assignee) Attorney
9. ☐ English Translation Document (if applicable)
10. ☒ Information Disclosure ☐ Copies of IDS
Statement (IDS)/PTO-1449 Citations
11. ☒ Preliminary Amendment
12. ☒ Return Receipt Postcard (MPEP 503)
(Should be specifically itemized)
13. ☐ * Small Entity ☐ Statement filed in prior application
Statement(s) Status still proper and desired
(PTO/SB/09-12)
14. ☐ Certified Copy of Priority Document(s)
(if foreign priority is claimed)
15. ☐ Other:

* NOTE FOR ITEMS 1 & 13 IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY
FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT
IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).

16. If a CONTINUING APPLICATION, check appropriate box, and supply the requisite information below and in a preliminary amendment:

☒ Continuation ☐ Divisional ☐ Continuation-in-part (CIP)

of prior application No: 09/429,398

Prior application information: Examiner J. Corriellus

Group / Art Unit: 2777

For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

17. CORRESPONDENCE ADDRESS

☒ Customer Number or Bar Code Label

020457

or ☐ Correspondence address below

(Insert Customer No. or Attach bar code label here)

Name

Address

City

State

Zip Code

Country

Telephone

Fax

Name (Print/Type)

Carl I. Brandidge

Registration No. (Attorney/Agent)

29,621

Signature

Date

9/19/2000

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

[illegible]

- [illegible]

FEE TRANSMITTAL for FY 2000

*Patent fees are subject to annual revision
Small Entity payments must be supported by a small entity statement,
otherwise large entity fees must be paid. See Forms PTO/SB/09-12
See 37 C.F.R. §§ 1.27 and 1.28.*

TOTAL AMOUNT OF PAYMENT (\$)**690.00**

Complete if Known

Application Number	
Filing Date	September 19, 2000
First Named Inventor	Masashi TSUCHIDA
Examiner Name	J. Corriellus
Group / Art Unit	2777
Attorney Docket No.	520.33330CC3

METHOD OF PAYMENT (check one)

1. ☐ The Commissioner is hereby authorized to charge indicated fees and credit any overpayments to:

Deposit Account Number

Deposit Account Name

☒ Charge Any Additional Fee Required Under 37 CFR §§ 1.16 and 1.17

2. ☒ Payment Enclosed:

☐ Check ☐ Money Order ☒ Other

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Code	Large Entity Fee (\$)	Small Entity Code	Small Entity Fee (\$)	Fee Description	Fee Paid
101	690	201	345	Utility filing fee	690.00
106	310	206	155	Design filing fee	
107	480	207	240	Plant filing fee	
108	690	208	345	Reissue filing fee	
114	150	214	75	Provisional filing fee	

SUBTOTAL (1) (\$)**690.00**

2. EXTRA CLAIM FEES

Total Claims	Extra Claims	Fee from below	Fee Paid
1	-20** = 0	18	0
1	-3** = 0	78	0
Multiple Dependent			0

**or number previously paid, if greater; For Reissues, see below

Large Entity Code	Large Entity Fee (\$)	Small Entity Code	Small Entity Fee (\$)	Fee Description
103	18	203	9	Claims in excess of 20
102	78	202	39	Independent claims in excess of 3
104	260	204	130	Multiple dependent claim, if not paid
109	78	209	39	** Reissue independent claims over original patent
110	18	210	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) (\$)**0.00**

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Code	Large Entity Fee (\$)	Small Entity Code	Small Entity Fee (\$)	Fee Description	Fee Paid
105	130	205	65	Surcharge - late filing fee or oath	0.00
127	50	227	25	Surcharge - late provisional filing fee or cover sheet	0.00
139	130	139	130	Non-English specification	0.00
147	2,520	147	2,520	For filing a request for reexamination	0.00
112	920*	112	920*	Requesting publication of SIR prior to Examiner action	0.00
113	1,840*	113	1,840*	Requesting publication of SIR after Examiner action	0.00
115	110	215	55	Extension for reply within first month	0.00
116	380	216	190	Extension for reply within second month	0.00
117	870	217	435	Extension for reply within third month	0.00
118	1,360	218	680	Extension for reply within fourth month	0.00
128	1,850	228	925	Extension for reply within fifth month	0.00
119	300	219	150	Notice of Appeal	0.00
120	300	220	150	Filing a brief in support of an appeal	0.00
121	260	221	130	Request for oral hearing	0.00
138	1,510	138	1,510	Petition to institute a public use proceeding	0.00
140	110	240	55	Petition to revive - unavoidable	0.00
141	1,210	241	605	Petition to revive - unintentional	0.00
142	1,210	242	605	Utility issue fee (or reissue)	0.00
143	430	243	215	Design issue fee	0.00
144	580	244	290	Plant issue fee	0.00
122	130	122	130	Petitions to the Commissioner	0.00
123	50	123	50	Petitions related to provisional applications	0.00
126	240	126	240	Submission of Information Disclosure Stmt	0.00
581	40	581	40	Recording each patent assignment per property (times number of properties)	0.00
146	690	246	345	Filing a submission after final rejection (37 CFR § 1.129(a))	0.00
149	690	249	345	For each additional invention to be examined (37 CFR § 1.129(b))	0.00
Other fee (specify) _____					0.00
Other fee (specify) _____					0.00

* Reduced by Basic Filing Fee Paid

SUBTOTAL (3) (\$)**0.00**

SUBMITTED BY

Name (Print/Type) Carl L. Brandidge	Registration No. (Attorney/Agent) 29,621	Telephone 703-312-6600
Signature	Date 9/19/2000	

WARNING:

Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Washington, DC 20231.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: M. TSUCHIDA, et al
Serial No.: Not yet assigned
Filed: September 19, 2000
For: METHOD AND SYSTEM OF DATABASE DIVISIONAL
MANAGEMENT FOR PARALLEL DATABASE SYSTEM
Group: 2777
Examiner: J. Corrielus

PRELIMINARY AMENDMENT

Assistant Commissioner for Patents
Washington, D.C. 20231

September 19, 2000

Sir:

The following amendments and remarks are respectfully submitted prior to the Rule 53(b) Continuation Application filed on even date.

IN THE SPECIFICATION

Please insert before the first line of the specification the following:

-- This is a continuation of application Serial No. 09/429,398, filed October 28, 1999; which is a continuation of Serial No. 09/153,612, filed September 15, 1998, now U.S. Patent No. 6,101,495; which is a divisional of application Serial No. 08/341,953, now U.S. Patent No. 5,813,005. --

Please amend the specification as follows:

Page 19, line 25, delete "IOS node 75" insert --IOS node 70--.

Page 57, line 24, delete "11" insert --21--.

IN THE CLAIMS

Please cancel claims 2-39 without prejudice or disclaimer of the subject matter thereof.

REMARKS

Entry of the above amendments prior to examination is respectfully requested.

Please charge any shortage in fees due in connection with the filing of this paper, or credit any overpayment of fees, to the deposit account of Antonelli, Terry, Stout & Kraus, LLP, Deposit Account No. 01-2135 (520.33330CC3).

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Carl I. Brundidge
Registration No. 29,621

CIB/jdc
(703) 312-6600

SPECIFICATION

TITLE OF THE INVENTION

Method and System of Database Divisional
5 Management for Parallel Database System

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a method and a
10 system of database divisional management for use with a
parallel database system. More particularly, the
invention relates to a database divisional management
method and a parallel database system whereby the number
of processors or the number of disk units for database
15 processing is optimized under given loads.

2. Description of the Prior Art

Parallel database systems are proposed
illustratively by David Dewitt and Jim Gray in "Parallel
Database Systems: The Future of High Performance
20 Database Systems" (CACM, Vol. 35, No. 6, 1992, pp. 85 -
98). Parallel database systems of the kind proposed
above involve having a plurality of processors tightly
or loosely connected to one another and subjected to
database divisional management.

25 How to configure a conventional parallel

database system is at the user's discretion. Once established, the conventional system configuration must remain as it is. With its lack of flexibility, the system configuration may be unsuitable from the start
5 for dealing with envisaged loads or may become incapable of addressing the loads some time after the start of its operative state. In such cases, the desired degree of parallel processing is not available and/or high-speed query processing is not implemented.

10

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a database divisional management method and a parallel database system whereby desired degrees of
15 parallel processing and high-speed query processing are available.

In carrying out the invention and according to a first aspect thereof, there is provided a database divisional management system for use with a parallel
20 database system having a storage medium, storage and management means for storing and managing a database in the storage medium, and a plurality of access means for accessing the database in response to query inputs. This database divisional management system comprises:
25 generation means for generating a procedure for

5

10

15

20

25

database in the storage medium, and a network for connecting the FES, BES and IOS nodes. This database divisional management method comprises the steps of: calculating the load pattern by which to perform

5 database processing using the processing procedure; and determining the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, the number of processors assigned to the IOS node, the number of disk units of the IOS node, and the
10 number of partitions of the disk units in accordance with the load pattern for data processing.

According to a third aspect of the invention, there is provided a database divisional management method for use with a parallel database system
15 comprising an FES node for analyzing and optimizing user queries and generating a processing procedure in response thereto, BES nodes having a storage medium (i.e., disk units) in which to store a database and capable of accessing the database on the basis of the
20 processing procedure generated by the FES node, and a network for connecting the FES and BES nodes. This database divisional management method comprises the steps of: calculating the load pattern by which to perform database processing using the processing
25 procedure; and determining the number of processors

assigned to the FES node, the number of processors
assigned to the BES nodes, the number of disk units of
the BES nodes, and the number of partitions of the disk
units in accordance with the load pattern for database
5 processing.

According to a fourth aspect of the invention,
there is provided a database divisional management
method for use with a parallel database system
comprising an FES node for analyzing and optimizing user
10 queries and generating a processing procedure in
response thereto, BES nodes for accessing a database on
the basis of the processing procedure generated by the
FES node, an IOS node having a storage medium (i.e.,
disk units) and capable of storing and managing the
15 database in the storage medium, and a network for
connecting the FES, BES and IOS nodes. This database
divisional management method comprises the steps of:
determining the upper limit number of pages which are
accessible in parallel and which require a constant time
20 each when the database is scanned for access thereto;
and determining the number of processors assigned to the
FES node, the number of processors assigned to the BES
nodes, the number of processors assigned to the IOS
node, the number of disk units of the IOS node, and the
25 number of partitions of the disk units in accordance

with the upper limit number of pages.

According to a fifth aspect of the invention, there is provided a database divisional management method for use with a parallel database system

5 comprising an FES node for analyzing and optimizing user queries and generating a processing procedure in response thereto, BES nodes having a storage medium (i.e., disk units) in which to store and manage a database and capable of accessing the database on the

10 basis of the processing procedure generated by the FES node, and a network for connecting the FES and BES nodes. This database divisional management method comprises the steps of: determining the upper limit number of pages which are accessible in parallel and

15 which require a constant time each when the database is scanned for access thereto; and determining the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, the number of disk units of the BES nodes, and the number of partitions of

20 the disk units in accordance with the upper limit number of pages.

According to a sixth aspect of the invention, there is provided a database divisional management method for use with a parallel database system

25 comprising an FES node for analyzing and optimizing user

queries and generating a processing procedure in response thereto, BES nodes for accessing a database on the basis of the processing procedure generated by the FES node, an IOS node having a storage medium (i.e., disk units) and capable of storing and managing the database in the storage medium, and a network for connecting the FES, BES and IOS nodes. This database divisional management method comprises the steps of: calculating the expected degree of parallelism p according to the load pattern based on the processing procedure; and determining the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, the number of processors assigned to the IOS node, the number of disk units of the IOS node, and the number of partitions of the disk units in accordance with the expected degree of parallelism p.

According to a seventh aspect of the invention, there is provided a database divisional management method for use with a parallel database system comprising an FES node for analyzing and optimizing user queries and generating a processing procedure in response thereto, BES nodes having a storage medium (i.e., disk units) in which to store and manage a database and capable of accessing the database on the

basis of the processing procedure generated by the FES node, and a network for connecting the FES and BES nodes. This database divisional management method comprises the steps of: calculating the expected degree
5 of parallelism p according to the load pattern based on the processing procedure; and determining the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, the number of disk units of the BES nodes, and the number of partitions of
10 the disk units in accordance with the expected degree of parallelism p .

In another preferred structure according to the invention, the database divisional management method further comprises the steps of: calculating the optimum
15 number of accessible pages m ; calculating the number of pages s ($= m/p$) in units of sub-key ranges if key range partitions exist; and having sub-key range partitions in units of s pages for inserting data into a disk apparatus.

20 According to an eighth aspect of the invention, there is provided a database divisional management method for use with a parallel database system comprising an FES node for analyzing and optimizing user queries and generating a processing procedure in
25 response thereto, BES nodes for accessing a database on

the basis of the processing procedure generated by the FES node, an IOS node having a storage medium (i.e., disk units) and capable of storing and managing the database in the storage medium, and a network for
5 connecting the FES, BES and IOS nodes. This database divisional management method comprises the steps of: detecting a load unbalance on the basis of such load information items as the number of accessed pages, the number of hit rows and the number of communications
10 acquired during execution of the processing procedure; and changing the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, the number of processors assigned to the IOS node, and the number of disk units of the IOS node so as
15 to eliminate the load unbalance.

According to a ninth aspect of the invention, there is provided a database divisional management method for use with a parallel database system comprising an FES node for analyzing and optimizing user
20 queries and generating a processing procedure in response thereto, BES nodes having a storage medium (i.e., disk units) in which to store and manage a database and capable of accessing the database on the basis of the processing procedure generated by the FES
25 node, and a network for connecting the FES and BES

nodes. The database divisional management method comprises the steps of: detecting a load unbalance on the basis of such load information items as the number of accessed pages, the number of hit rows and the number of communications acquired during execution of the processing procedure; and changing the number of processors assigned to the FES node, the number of processors assigned to the BES nodes, and the number of disk units of the BES nodes so as to eliminate the load unbalance.

In a further preferred structure according to the invention, the database divisional management method further comprises the steps of: closing, when online processing is in progress, the key range of a database table if at least one of the three numbers consisting of the number of processors assigned to the BES nodes, the number of processors assigned to the IOS node and the number of disk units is to be increased, the database table being the object to be managed by either the processors or the disk units to be added; assigning the processors and the disk units anew; succeeding lock information and directory information; updating the dictionary information necessary for node assignment control; and releasing the closing of the key range thereafter if the online processing is still in

progress.

In an even further preferred structure according to the invention, the database divisional management method further comprises the steps of:

- 5 closing, when online processing is in progress, the key range of a database table if either the number of processors assigned to the BES nodes or the number of disk units is to be increased, the database table being the object to be managed by either the processors or the
- 10 disk units to be added; assigning either the processors or the disk units anew; succeeding lock information and directory information; updating the dictionary information necessary for node assignment control; moving data from the existing group of disk units to the
- 15 newly added disk units; and releasing the closing of the key range thereafter if the online processing is still in progress.

In a still further preferred structure according to the invention, the database divisional

- 20 management method comprises the steps of: closing, when online processing is in progress, the key range of a database table if at least one of the three numbers consisting of the number of processors assigned to the BES nodes, the number of processors assigned to the IOS
- 25 node and the number of disk units is to be decreased,

the database table being managed by either the processors or the disk units to be removed; determining either the processors or the disk units to be removed; succeeding lock information and directory information; 5 updating the dictionary information necessary for node assignment control; and releasing the closing of the key range thereafter if the online processing is still in progress.

In another preferred structure according to the invention, the database divisional management method further comprises the steps of: closing, when online processing is in progress, the key range of a database table if at least either the number of processors assigned to the BES nodes or the number of disk units is to be decreased, the database table being managed by either the processors or the disk units to be removed; determining either the processors or the disk units to be removed; succeeding lock information and directory information; updating the dictionary information necessary for node assignment control; moving data from the disk units to be removed to the disk units succeeding those units to be removed; and releasing the closing of the key range thereafter if the online processing is still in progress.

25 In a further preferred structure according to

the invention, either the number of processors or the number of disk units for database processing is changed dynamically.

In operation, the invention of the aspects and preferred structures outlined above provides the following major functions and features:

The database divisional management method according to the second aspect of the invention determines the number of processors assigned to each of the configured nodes, the number of disk units, and the number of partitions of the disk units in accordance with the load pattern for database processing (single item search, single item update, data retrieval, etc.). The invention embodied in this structure provides a system configuration suitable for dealing with the load in question, offers the expected degree of parallelism and permits high-speed query processing.

The database divisional management method according to the fourth aspect of the invention determines the number of processors assigned to each of the configured nodes, the number of disk units, and the number of partitions of the disk units in accordance with the upper limit number of pages which are accessible in parallel and which require a constant time each when the database is scanned for access thereto.

The invention embodied in this structure realizes high-speed query processing.

The database divisional management method according to the sixth aspect of the invention
5 determines the number of processors assigned to each of the configured nodes, the number of disk units, and the number of partitions of the disk units in accordance with the expected degree of parallelism p according to the load pattern. The invention embodied in this
10 structure provides the expected degree of parallelism.

The database divisional management method of a preferred structure calculates the number of pages s in units of sub-key ranges using the optimum number of accessible pages m and the expected degree of
15 parallelism p ($s = m/p$), and gets sub-key range partitions in units of s pages for inserting data into a disk apparatus. The invention embodied in this alternative structure allows data to be managed in substantially equal partitions.

20 The database divisional management method according to the eighth aspect of the invention detects a load unbalance, and changes the number of processors assigned to each of the configured nodes or the number of disk units so as to eliminate the detected load
25 unbalance. The invention embodied in this structure

keeps the system always configured to allow for the load fluctuation, provides the expected degree of parallelism and permits high-speed query processing.

When online processing is in progress, the
5 database divisional management method of another
alternative structure closes the key range of the
relevant database table if either the number of
processors assigned to each of the configured nodes or
the number of disk units is to be increased or
10 decreased. After this, necessary information is
succeeded and the closing of the key range is released.
This structure minimizes the overhead and, in a system
configuration comprising the IOS node, allows the data
to be succeeded without moving them across the disk
15 units.

The database divisional management method of
yet another preferred structure allows either the number
of processors or the number of disk units for database
processing to be changed dynamically. This feature
20 readily provides for a scalable parallel database
system.

The foregoing and other objects, advantages,
manner of operation and novel features of the present
invention will be understood from the following detailed
25 description when read in connection with the

accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a parallel
5 database system embodying the invention;

Fig. 2 is a conceptual view of a database
divisional management method embodying the invention;

Fig. 3 is a conceptual view of optimal node
distribution (where an IOS exists) by the inventive
10 database divisional management method;

Fig. 4 is a conceptual view of optimal node
distribution (where an IOS does not exist) by the
inventive database divisional management method;

Fig. 5 is a block diagram of an FES;
15

Fig. 6 is a block diagram of a BES;

Fig. 7 is a block diagram of an IOS;

Fig. 8 is a block diagram of a DS;

Fig. 9 is a block diagram of a JS;

Fig. 10 is a flowchart of steps performed by a
20 system controller;

Fig. 11 is a flowchart of steps representing a
query analysis process;

Fig. 12 is a flowchart of steps representing
query analysis;

25 Fig. 13 is a flowchart of steps representing

006760 8119960

```
static optimization processing;
```

Fig. 14 is a flowchart of steps representing predicate selectivity estimation processing;

Fig. 15 is a flowchart of steps representing
5 access path pruning;

Fig. 16 is a flowchart of steps representing processing procedure candidate generation;

Fig. 17 is a flowchart of steps representing code generation processing;

10 Fig. 18 is a flowchart of steps representing
query execution processing;

Fig. 19 is a flowchart of steps representing dynamic optimization processing;

Fig. 20 is a flowchart of steps representing
15 code interpretation processing;

Fig. 21 is a flowchart of steps representing data load processing;

Fig. 22 is a flowchart of steps representing dynamic load control processing; and

20 Fig. 23 is a conceptual view of dynamic load
control.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Preferred embodiments of the invention will now
25 be described with reference to the accompanying

drawings. It should be noted that such description is for illustrative purposes only and is not limitative of the invention.

Fig. 1 is a block diagram of a parallel database system 1 embodying the invention. The parallel database system 1 comprises an FES (front end server) node, a BES (back end server) node, an IOS (input/output server) node, a DS (dictionary server) node and a JS (journal server) node, all connected by a network 90. Each of the configured nodes is also connected to another system.

The FES node 75 is composed of at least one processor having no disk units. This node has front end server functions for analyzing and optimizing user queries and for generating a processing procedure in response thereto.

The BES node 73 comprises at least one processor with no disk units. This node is capable of accessing a database by use of the processing procedure generated by the FES 75.

The IOS node 70 includes at least one processor and is equipped with at least one disk unit 80. The disk unit 80 accommodates the database so that the latter may be managed therein.

If the BES node 73 is supplemented by the

functions of the IOS nodes, the IOS node may be omitted. In such a case, the disk unit 80 is connected to the BES node 73 which now has functions for storing and managing the database in the connected disk unit 80.

5 The database is composed of a plurality of tables. Each table is a two-dimensional table comprising a plurality of rows. One row includes at least one column (i.e., attribute). These tables, stored in the disk unit 80, are each divided physically
10 into fixed-length pages made of a predetermined number of rows each. The location at which each of the pages is stored in the disk unit 80 is known by referencing directory information.

 The DS node 71 comprises at least one processor
15 and has at least one disk unit 81. This node is capable of managing database definition information as a whole.

 The JS node 72 includes at least one processor and has at least one disk unit 82. This node has
20 functions for storing and managing the history information on database updates carried out by each node.

 Fig. 2 is a conceptual view of a database divisional management method embodying the invention. This method determines the number of processors assigned
25 to each of the FES node 75, BES node 73 and IOS node 75

in Fig. 1, the number of configured disk units, and the number of partitions of the disk units.

First to be determined is the load pattern of database processing on the basis of the use-designated work load. The load pattern is any one of a diverse kinds of processing including single item search, single item update and data retrieval. With the load pattern established, the IOS node 70 determines accordingly the number of partitions of the disk units 80 for management. (If the BES node 73 incorporates IOS node functions, the BES node 73 determines the number of partitions of the disk units for management.)

Specifically, at the time of schema definition, the number of necessary disk units is determined by the way the database tables are partitioned (in terms of key range, number of rows per range converted from the number of pages, etc.). When the unit in which the key range is closed or reconfigured is determined, the combination of BES and IOS units 73 and 70 is determined (the unit in which to close or reconfigure the key range is dependent on the configuration made of disk units and IOS and BES nodes).

In the manner described, the configuration of the BES and IOS units 73 and 70 and of the disk units 80 is determined as follows:

- Partition count: unit number of database partitions that may be accessed in parallel under management of all BES nodes

- Disk count per partition: number of disk units
5 assigned to each partition

Fig. 2 shows an example in which the disk count is 8, the partition count is 4 and the disk count per partition is 2.

If the processor performance is enhanced by a
10 factor of n , then the number of volumes for use by each partition is multiplied by n . (It should be noted that the number of disk units is limited because of the constraints on the overall data transfer rate between the IOS node 70 and the disk units 80.)

15 Although one disk unit represents one disk drive in the above example, the one-to-one correspondence between a disk drive and a disk unit is not mandatory for this invention. One disk unit may illustratively contain a plurality of disk drives (e.g.,
20 disk array apparatus). In that case, the unit number of I/O units that may be accessed in parallel may be regarded as a disk unit.

In the example of Fig. 2 where the normal configuration is composed of one FES node, four BES
25 nodes, one IOS node and eight disk units, there need

only be one FES node and one BES node at initial data load time. That is, the initial configuration is made of one FES node, one BES node, one IOS node and eight disk units. Thus a BES node 731 has directory
5 information about the database stored in disk units 811 - 842 constituting partitions #1 - #4.

Where the BES load is so low that one BES node 731 is enough to deal with the IOS 70 and eight disk units 811 - 842, the single BES node 731 alone accesses
10 the database stored in the eight disk units 811 - 842. In that case, the configuration remains composed of one FES node, one BES node, one IOS node and eight disk units.

If the load on the BES node 731 is on the
15 increase and its activity ratio stays at 100%, a load unbalance may eventually be detected. In that case, another BES node 732 is added to the configuration. Since the partition count is 4, the two BES nodes 731 and 732 are assigned two partitions each. The BES node
20 731 thus has directory information about the database stored in the disk units 811 - 822 constituting partitions #1 - #2; the BES node 732 has directory information about the database stored in the disk units 831 - 842 constituting partitions #3 - #4. The
25 resulting configuration is composed of one FES node, two

BES nodes, one IOS node and eight disk units.

5 If the load on the BES nodes 731 and 732 is
still on the increase and their activity ratio stays at
100%, a load unbalance may also be detected. In that
10 case, the BES nodes 731 and 732 are supplemented
respectively by BES nodes 733 and 734. Since the
partition count is 4, the four BES nodes 731, 732, 733
and 734 are assigned one partition each. The BES node
731 thus has directory information about the database
15 stored in the disk units 811 - 812 constituting
partition #1; the BES node 732 has directory information
about the database stored in the disk units 821 - 822
constituting partition #2; the BES node 733 has
directory information about the database stored in the
15 disk units 831 - 832 constituting partition #3; and the
BES node 734 has directory information about the
database stored in the disk units 841 - 842 constituting
partition #4. The resulting configuration is composed
of one FES node, four BES nodes, one IOS node and eight
20 disk units.

When the load is on the decrease and the
activity ratio of the BES nodes 733 and 734 drops
illustratively below the 50% benchmark and remains
there, it then becomes more efficient for the BES nodes
25 733 and 734 to be assigned to other tasks. The BES

nodes 733 and 734 whose activity ratio is below 50% are thus reconfigured. The reduced configuration comprises one FES node, two BES nodes, one IOS node and eight disk units.

5 As described, where the number of BES nodes is increased or reduced depending on the amount of load, a scalable system is implemented within a range defined by two scales of configuration: one FES node, one BES node, one IOS node and eight disk units on the one hand; and
10 one FES node, four BES nodes, one IOS node and eight disk units on the other.

The IOS node 70 need only address parallel tasks as many as parallelly accessible disk units regardless of the correspondence between the BES nodes 73 and the disk units 80. This makes it possible to change the correspondence between the BES nodes 73 and the disk units 80 by moving the directory information across the BES nodes without recourse to data movement. Separating and integrating accesses is thus made easier.

20 Below is a description of how two
representative load patterns of single item update and
data retrieval are processed. The processing is
described in terms of the number of steps involved, the
number of processors, the number of disk units and the
25 number of partitions of the disk units. The

preconditions for the processing are assumed as follows:

FES processing (data reception): 30 (K steps)
BES processing (single item update): 60 (K steps)
BES processing (data retrieval): 220 (K steps)
5 Transmission: 6 (K steps)
Reception: $6 + 4 * \text{page count}$ (K steps)
Issuance of input/output request:
 $4 + 4 * \text{page count}$ (K steps)
Processor performance: 10 (M steps per second)
10 Input/output performance (one-page access): 20 (msec)
Input/output performance (10-page access): 30 (msec)

A. Single item update (one-page access)

(1) With system configuration comprising IOS node
15 Dividing the processor performance (10 M steps
per second) by the FES processing step count (30 K
steps) provides an available reception count of up to
333 times per second.

The necessary step count for single item update
20 by a BES node is given as: reception of an execution
request from the FES node (6 K steps) + transmission of
a data retrieval request from the BES node (6 K steps)
+ reception of retrieved data from IOS node (10 K steps)
+ single item update (60 K steps) + transmission of the
25 result of the execution request to FES node (6 K steps)

= 88 (K steps). Dividing the processor performance (10 M steps per second) by the obtained single item update step count (88 K steps) for the BES node provides an available single item update count of up to 114 times per second.

In addition, the necessary step count for the IOS node to access disk units is given as: reception of an input/output request from the BES node (6 K steps) + issuance of the input/output request (8 K steps) + transmission of the result of the input/output request to the BES node (6 K steps) = 20 (K steps). Dividing the processor performance (10 M steps per second) by the obtained disk access step count of 20 (K steps) for the IOS node provides an available disk access count of up to 500 times per second.

Because it takes 20 msec to perform random input/output of one page, one disk unit may be accessed up to 50 times per second. When the maximum available disk access count of 500 times per second for the IOS node is divided by the single disk access count of 50 times per second, the result indicates that up to 10 disk units may be connected to the IOS node.

When the maximum available disk access count of 500 times per second for the IOS node is divided by the single item update count of 114 times per second for the

BES node, the result indicates that one IOS node may address up to 4.3 BES nodes.

When the maximum available reception count of up to 333 times per second for the FES node is divided
5 by the single item update count of 114 times per second for the BES node, the result indicates that one FES node may address up to three BES nodes.

The results above are summarized into ratios of
FES : BES = 1 : 3, BES : IOS = 4.3 : 1, and IOS : disk
10 units = 1 : 10. Putting these ratios together provides a substantially balanced configuration composed of one FES node, four BES nodes, one IOS node and eight disk units, as shown in Fig. 3 (with a minor imbalance regarding the FES node and the disk units).

15 (2) With system configuration where BES nodes furnish IOS node functions

Dividing the processor performance (10 M steps per second) by the FES processing step count (30 K steps) provides an available reception count of up to
20 333 times per second.

The necessary step count for single item update by a BES node is given as: reception of an execution request from the FES node (6 K steps) + issuance of an input/output request (8 K steps) + single item update
25 (60 K steps) + transmission of the result of the

execution request to FES node (6 K steps) = 80 (K steps). Dividing the processor performance (10 M steps per second) by the obtained single item update step count (80 K steps) provides an available single item
5 update count of up to 125 times per second.

Because it takes 20 msec to perform random input/output of one page, one disk unit may be accessed up to 50 times per second. When the maximum available single item update count of 125 times per second for the
10 BES node is divided by the single disk access count of 50 times per second, the result indicates that up to 2.5 disk units may be connected to the BES node.

When the maximum available reception count of up to 333 times per second for the FES node is divided
15 by the single item update count of 125 times per second for the BES node, the result indicates that one FES node may address up to 2.6 BES nodes.

The results above are summarized into ratios of
FES : BES = 1 : 2.6 and BES : disk units = 1 : 2.5.
20 Putting these ratios together provides a substantially balanced configuration composed of one FES node, four BES nodes and eight disk units, as shown in Fig. 4 (with a minor imbalance regarding the FES node).

25 B. Data retrieval (10-page access)

(1) With system configuration comprising IOS node

Dividing the processor performance (10 M steps per second) by the FES processing step count (30 K steps) provides an available reception count of up to

5 333 times per second.

The necessary step count for data retrieval by a BES node is given as: reception of an execution request from the FES node (6 K steps) + transmission of a data retrieval request from the BES node (6 K steps)
10 + reception of retrieved data from IOS node (46 K steps) + retrieval of data (220 K steps) + transmission of the result of the execution request to FES node (6 K steps) = 284 (K steps). Dividing the processor performance (10 M steps per second) by the obtained data retrieval step
15 count (284 K steps) provides an available data retrieval count of up to 35 times per second.

In addition, the necessary step count for the IOS node to access disk units is given as: reception of an input/output request from the BES node (6 K steps) +
20 issuance of the input/output request (44 K steps) + transmission of the result of the input/output request to the BES node (6 K steps) = 56 (K steps). Dividing the processor performance (10 M steps per second) by the obtained step count provides an available disk access
25 count of up to 179 times per second.

Because it takes 30 msec to perform batch
input/output of 10 pages, one disk unit may be accessed
up to 33 times per second. When the maximum available
disk access count of 179 times per second for the IOS
5 node is divided by the single disk access count of 33
times per second, the result indicates that up to 5.4
disk units may be connected to the IOS node.

When the maximum available disk access count of
179 times per second for the IOS node is divided by the
10 available data retrieval count of 35 times per second
for the BES node, the result indicates that one IOS node
may address up to 5.1 BES nodes.

In addition, when the maximum available
reception count of 333 times per second for the FES node
15 is divided by the available data retrieval count of 35
times per second for the BES node, the result indicates
that one FES node may address up to 9.5 BES nodes.

The results above are summarized into ratios of
FES : BES = 1 : 9.5, BES : IOS = 5.1 : 1, and IOS : disk
20 units = 1 : 5.4. Putting these ratios together provides
a substantially balanced configuration composed of one
FES node, 10 BES nodes, two IOS nodes and 10 disk units
(with a minor imbalance regarding the disk units).

(2) With system configuration where BES nodes
25 furnish IOS node functions

Dividing the processor performance (10 M steps per second) by the FES processing step count (30 K steps) provides an available reception count of up to 333 times per second.

5 The necessary step count for data retrieval by
a BES node is given as: reception of an execution
request from the FES node (6 K steps) + issuance of an
input/output request (44 K steps) + retrieval of data
(220 K steps) + transmission of the result of the
10 execution request to FES node (6 K steps) = 276 (K
steps). Dividing the processor performance (10 M steps
per second) by the obtained data retrieval step count
(276 K steps) for the BES node provides an available
data retrieval count of up to 36 times per second.

15 Because it takes 30 msec to perform batch
input/output of 10 pages, one disk unit may be accessed
up to 33 times per second. When the maximum available
data retrieval count of 36 times per second for the BES
node is divided by the single disk access count of 33
20 times per second, the result indicates that one disk
unit may be connected to the BES node.

When the maximum available reception count of up to 333 times per second for the FES node is divided by the available data retrieval count of 36 times per second for the BES node, the result indicates that one

FES node may address up to 9.2 BES nodes.

The results above are summarized into ratios of

FES : BES = 1 : 9.2 and BES : disk unit = 1 : 1.1

Putting these ratios together provides a substantially
5 balanced configuration composed of one FES node, 10 BES
nodes and 10 disk units.

Fig. 5 is a block diagram of the FES node 75. The FES node 75 comprises user-generated application programs 10 - 11, a parallel database management system 20 for providing overall database system management such as query processing and resource management, and an operating system 30 for managing all computer system operations including the reading and writing of data.

The parallel database management system 20 has a system controller 21, a logical processor 22, a physical processor 23, and a database dictionary buffer 24 for temporarily accommodating target data. The parallel database management system 20 is connected to the network 90 as well as to another parallel database management system.

The system controller 21 primarily manages input/output and other operations. The system controller 21 has a data load processing program 210 and a dynamic load control processing program 211.

25 The logical processor 22 includes a query

analysis program 220 for analyzing queries in terms of
syntax and semantics, a static optimization processing
program 221 for generating appropriate processing
procedure candidates, and a code generator 222 for
5 generating the code applicable to each processing
procedure candidate generated. The logical processor 22
further includes a dynamic optimization processing
program 223 for selecting an optimal processing
procedure and a code interpreter 224 for interpreting
10 the code of the selected processing procedure candidate.

The physical processor 23 comprises a data
access processing program 230 that edits accessed data
and judges them for conditions, and adds records; a
database dictionary buffer controller 231 that controls
15 the reading and writing of database records; and a
concurrency controller 233 that provides concurrency
control over the resources shared by the systems.

Fig. 6 is a block diagram of the BES node 73.
The BES node 73 is composed of the parallel database
20 management system 20 for overall database system
management and the operating system 30 for overall
computer system management. Where the BES node 73 is
equipped with IOS node functions, the BES node 73 has
disk units in which to store and manage a database 40.

25 The parallel database management system 20

5

10

```
interpreter 224 that carries out code interpretation.
```

15

25

system 30 for overall computer system management. The disk unit 80 contains the database 40.

The parallel database management system 20 comprises the system controller 21, the physical processor 23, and an input/output buffer 24 for temporarily accommodating target data. This parallel database management system 20 is connected to the network 90 as well as to another parallel database management system.

The system controller 21 manages input/output and other operations. The system controller 21 includes the data load processing program 210 that loads data by taking load distribution into consideration.

The physical processor 23 comprises the data access processing program 230 that edits accessed data and judges them for conditions, and adds records; and an input/output buffer controller 231 that controls the reading and writing of database records.

Fig. 8 is a block diagram of the DS 71 together with the disk unit 81. The DS 71 comprises the parallel database management system 20 for overall database system management and the operating system 30 for overall computer system management. The disk unit 81 contains a dictionary 50.

The parallel database management system 20

comprises the system controller 21, the logical processor 22, the physical processor 23, and a dictionary buffer 24. This parallel database management system 20 is connected to the network 90 as well as to
5 another parallel database management system.

The logical processor 22 includes the code interpreter 224 that carries out code interpretation.

The physical processor 23 includes the data access processing program 230 that edits accessed data
10 and judges them for conditions, and adds records; the dictionary buffer controller 231 that controls the reading and writing of dictionary records; and the concurrency controller 233 that provides concurrency control over the resources shared by the systems.

15 Fig. 9 is a block diagram of the JS 72 together with the disk unit 82. The JS 72 comprises the parallel database management system 20 for overall database system management and the operating system 30 for overall computer system management. The disk unit 82
20 contains a journal 60.

The parallel database management system 20 comprises the system controller 21, the physical processor 23, and a journal buffer 24. This parallel database management system 20 is connected to the
25 network 90 as well as to another parallel database

management system.

The physical processor 23 includes the data access processing program 230 that edits accessed data and judges them for conditions, and adds records; and a
5 journal buffer controller 231 that controls the reading and writing of journal records.

Fig. 10 is a flowchart of steps performed by the system controller of the parallel database management system 20 in the FES node 75. The system
10 controller 21 first checks to see if a query analysis process is selected (step 212). If the query analysis process is found to be selected, the system controller 21 calls and executes a call query analysis program 400. After execution of the call query analysis program 400,
15 the system controller 21 ends its operation.

If the query analysis process is not in effect in step 212, the system controller 21 checks to see if a query execution process is selected (step 213). If the query execution process is found to be selected, the
20 system controller 21 calls and executes a query execution program 410. After execution of the query execution program 410, the system controller 21 ends its operation.

If the query execution process is not in effect
25 in step 213, the system controller 21 checks to see if a

data load process is selected (step 214). If the data load process is found to be selected, the system controller 21 calls and executes a data load program 210. After execution of the data load program 210, the system controller 21 ends its operation.

If the data load process is not in effect in step 214, the system controller 21 checks to see if a dynamic load control process is selected (step 215). If the dynamic load control program is found to be selected, the system controller 21 calls and executes a dynamic load control program 211. After execution of the dynamic load control program 211, the system controller 21 end its operation.

If the dynamic load control process is not in effect in step 215, then the system controller 21 terminates its operation.

The flowchart of steps performed by the database management system 20 of the BES node 73 is that of Fig. 10 minus steps 212, 215, 400 and 211. The flowchart of steps carried out by the database management system 20 of the IOS node 70 is that of Fig. 10 minus steps 212, 213, 215, 400, 410 and 211.

Fig. 11 is a flowchart of steps performed by the query analysis program 400. In step 220 for query analysis, the program 400 analyzes the input query

sentence for syntax and semantics.

In step 221 for static optimization, the query analysis program 400 estimates the ratio of the data that would satisfy the condition occurring in the query.

5 Then based on predetermined rules, the program 400 generates effective access path candidates (especially indices) so as to prepare a processing procedure candidate.

10 In step 222 for code generation, the query analysis program 400 translates the processing procedure candidate into an executable form code. This terminates the processing of the query analysis program 400.

15 Fig. 12 is a flowchart of steps representing the query analysis 220. In step 2200, the input query sentence is analyzed for syntax and semantics. Then the current processing is terminated.

20 Fig. 13 is a flowchart of steps representing the static optimization process 221. In step 2210 for predicate selectivity estimation, the selectivity of the predicate of the condition occurring in the query is estimated.

In step 2211 for access path pruning, the access paths including indices are pruned.

25 In step 2212 for processing procedure candidate generation, the processing procedure candidate combining

the access paths is generated. This terminates the static optimization processing.

Fig. 14 is a flowchart of steps representing the predicate selectivity estimation process 2210. In
5 step 22101, a check is made to see if any variable appears in the query condition. If no variable appears in the query condition, step 22102 is reached. If a variable appears in the query condition, step 22104 is reached.

10 In step 22102, a check is made to see if the query condition contains column value frequency information. If the column value frequency information is present, step 22103 is reached. If the column value frequency information is not found, step 22105 is
15 reached.

In step 22103, the selectivity is calculated by use of the column value frequency information, and the current processing is terminated.

20 In step 22104, a check is made to see if the query condition contains column value frequency information. If the column value frequency information is present, the current processing is terminated; if no such information exists, step 22105 is reached.

25 In step 22105, a default value is set in accordance with the kind of the query condition. The

006160 344960

current processing is then terminated.

Fig. 15 is a flowchart of steps representing the access path pruning process 2212. In step 22120, column indices appearing in the condition are set as
5 access path candidates.

In step 22121, a check is made to see if the table to be accessed for the query is stored separately in a plurality of nodes. If the table is not stored separately, step 22122 is reached; if the table is
10 stored separately, step 22123 is reached.

In step 22122, sequential scans are set as access path candidates.

In step 22123, parallel scans are set as access path candidates.

15 In step 22124, a check is made to see if the selectivity of each condition is already set. If the selectivity is found to be already set, step 22125 is reached; if the selectivity has yet to be set, step 22126 is reached.

20 In step 22125, the highest access path priority is given to the index of the condition whose selectivity is the smallest regarding each table.

In step 22126, the maximum and minimum values of the selectivity of each condition are obtained.

25 In step 22127, the reference for selecting each

access path is calculated from system characteristics such as the processor performance and I/O performance.

In step 22128, only those access paths of combined single or plural indices whose selectivity is less than the above reference are set as access path candidates.

Fig. 16 is a flowchart of steps representing the processing procedure candidate generation process 2213. In step 22130, a check is made to see if the table to be accessed for the query is stored separately in a plurality of nodes. If the table is not stored separately, step 22131 is reached; if the table is stored separately, step 22135 is reached.

In step 22131, a check is made to see if a sorting process is included in the processing procedure candidate. If the sorting process is not included, step 22132 is reached; if the sorting process is included, step 22135 is reached.

In step 22132, a check is made to see if there is only one access path of the table to be accessed. If only one access path exists, step 22133 is reached; if more than one access path is present, step 22134 is reached.

In step 22133, a single processing candidate is generated, and the current processing is terminated.

In step 22134, a plurality of processing candidates are generated, and the current processing is terminated.

In step 22135, the query is broken up into joinable two-way joins.

In step 22136, data read/data distribution processing procedures and slot sort processing procedures are set as candidates in correspondence with the stored nodes of the table stored separately.

In step 22137, slot sort processing procedures, N-way merge processing procedures and join processing procedures are set as candidates in correspondence with the join processing nodes. The slot sort processing refers to an intra-page sorting process in which the page for accommodating data has each of its rows managed by a slot that is offset from the beginning of the page. When the slots are read consecutively, the corresponding rows are accessed in ascending order. The N-way merge processing refers to a process in which an N-way buffer is used to input N runs at each merge stage for sorting by tournament into a single sort run.

In step 22138, a requested data output processing procedure is set to the request data output node.

In step 22139, a check is made to see if

estimations are finished on all tables. If not all estimations are finished, step 22136 is reached again; if all estimations are found to be finished, the current processing is terminated.

5 Fig. 17 is a flowchart of steps carried out by
the code generator 222. In step 2220, a check is made
to see if there is only one processing procedure
candidate. If more than one processing procedure
candidate is found, step 2221 is reached; if there
0 exists only one processing procedure candidate, step
2223 is reached.

In step 2221, optimization information composed of column value frequency information and others is embedded into the processing procedures.

15 In step 2222, the data structure for selecting
processing procedures is generated on the basis of the
constants substituted upon query execution.

In step 2223, the processing procedure is translated into execution formulae. The processing is then terminated.

Fig. 18 is a flowchart of steps performed by the query execution program 410. In step 223 for dynamic optimization processing, the processing procedure to be executed on each node group is determined on the basis of the substituted constants.

In step 224 for code interpretation, the processing procedure is interpreted and executed. The current processing is then terminated.

Fig. 19 is a flowchart of steps performed by the dynamic optimization processing program 223. In step 22300, a dynamic load control program is called and executed.

In step 22301, a check is made to see if there is only one processing procedure being generated. If there is only one processing procedure being generated, the current processing is terminated. If there is more than one processing procedure being generated, step 22302 is reached.

In step 22302, the selectivity is calculated on the basis of the substituted constants.

In step 22303, a check is made to see if parallel processing procedure candidates are included in the processing procedure candidates. If parallel processing procedure candidates are found to be included, step 22304 is reached; if no such candidates are found, step 22308 is reached.

In step 22304, the optimization information (column value frequency information of join columns, number of rows in the table to be accessed, page count, etc.) is input from the dictionary.

In step 22305, the processing time for data retrieval and data distribution is calculated in consideration of various system characteristics.

In step 22306, the assigning number p of join
5 nodes is determined based on the processing time.

In step 22307, data distribution information is generated on the basis of the optimization information.

In step 22308, the processing procedure is selected by use of the access path selection reference.

10 Fig. 20 is a flowchart of steps carried out by the code interpreter 224. In step 22400, a check is made to see if data retrieval and a data distribution process are in effect. If data retrieval and the data distribution process are found to be in effect, step
15 22401 is reached. If data retrieval and the data distribution process are not found to be in effect, step 22405 is reached.

In step 22401, the database is accessed and the condition is evaluated.

20 In step 22402, data is set into a buffer of each node based on the data distribution information.

In step 22403, a check is made to see if the buffer of the node in question is full. If the buffer is found to be full, step 22404 is reached; if the
25 buffer is not full, step 22420 is reached.

In step 22404, data is transferred in page form to the corresponding node. Step 22404 is followed by step 22420.

5 In step 22405, a check is made to see if a slot sort process is in effect. If the slot sort process is found to be in effect, step 22406 is reached; if the slot sort process is not found to be in effect, step 22409 is reached.

10 In step 22406, page form data is received from another node.

In step 22407, the slot sort process is executed.

15 In step 22408, the result of the slot sort process is stored temporarily. Step 22408 is followed by step 22420.

20 In step 22409, a check is made to see if an N-way merge process is in effect. If the N-way merge process is found to be in effect, step 22410 is reached; if the N-way merge process is not found to be in effect, step 22412 is reached.

In step 22410, the N-way merge process is executed.

25 In step 22411, the result of the N-way merge process is stored temporarily. Step 22411 is followed by step 22420.

In step 22412, a check is made to see if a join process is in effect. If the join process is found to be in effect, step 22413 is reached; if the join process is not found to be in effect, step 22416 is reached.

5 In step 22413, both sort lists are joined and
data is set to a buffer for output.

In step 22414, a check is made to see if the buffer for output is full. If the buffer for output is found to be full, step 22415 is reached; if the buffer
10 for output is not found to be full, step 22420 is reached.

In step 22415, data is transferred in page form to the request data output node. Step 22415 is followed by step 22420.

15 In step 22416, a check is made to see if a
request data output process is in effect. If the
request data output process is found to be in effect,
step 22417 is reached. If the request data output
process is not found to be in effect, step 22420 is
20 reached.

In step 22417, a check is made to see if page form data is transferred from another node. If page form data is found to be transferred from another node, step 22418 is reached; if no such data is found to be transferred, step 22419 is reached.

In step 22418, page form data is received from another node.

In step 22419, the result of the query processing is output to the application program.

5 In step 22420, a check is made to see if the
BES node is in execution. If the BES node is found to
be in execution, step 22421 is reached; if the BES node
is not in execution, the current processing is
terminated.

10 In step 22421, the FES is notified of that
information for estimating processing load which
includes the access page count, the hit row count and
the number of communications. The current processing is
then terminated.

Fig. 21 is a flowchart of steps performed by the data load program 210. Before each step of the data load process is explained, general concepts of the process will be outlined below. The data load method comes in three kinds: one method distributes data with the emphasis on a target response time, whereby the time required to scan the entire table is limited below a predetermined level; another method distributes data with the emphasis on degrees of parallelism, whereby m pages are accessed for optimal parallel processing; yet another method distributes data as desired under user

The distribution of data with the emphasis on the target response time involves initially finding the number of pages in which to store the rows of the entire table. Then the upper limit number of pages to be stored in the disk units partitioned for parallel access is determined. If necessary, batch input (e.g., of 10 pages) is presupposed for access. Load distribution is determined in view of the combination of the number of disk units, the number of IOS nodes and the number of BES nodes. If there exists a key range division, the divided key range is subdivided by the upper limit page count and data is stored into the subdivided key ranges of the disk units. More aspects of this process will be described later in detail with reference to Fig. 23.

25 The expected degree of parallelism p is

calculated as the square root of the ratio given by
dividing the response time by the overhead per node.
When so obtained, the ratio of 10 corresponds to the
expected degree of parallelism of 3, 100 to 10, 1,000 to
5 32, and 10,000 to 100. If the calculated degree of
parallelism p is higher than the existing partition
count, the existing partition count is selected (because
it determines the maximum number of disk units that may
be processed). Conversely, if the calculated expected
10 degree of parallelism p is lower than the existing
partition count, the expected degree of parallelism p is
selected with the existing partition count taken as the
upper limit.

Specifically, suppose that the distribution of
15 data optimized for 100-page access is to be calculated.
As a precondition, batch input is to be carried out in
units of 10 pages. Because it takes 300 msec to perform
one I/O operation (i.e., 10-page access) and 5.6 msec to
carry out one I/O overhead operation (56 ks required for
20 10-MIPS performance), the expected degree of parallelism
 p is about 7 ($=\sqrt{300/5.6}$). Thus the key range is
subdivided into sub-key ranges in units of 14 pages (s
 $= 100/7$).

The user-designated distribution of data is
25 effected in the same manner as with conventional

database management systems. That is, data distribution is managed as set by parameters.

In step 21000 of Fig. 21, a check is made to see if data distribution puts the emphasis on the target response time. If the emphasis is not found to be placed on the target response time, step 21001 is reached. If the emphasis is found to be placed on the target response time, step 21003 is reached.

In step 21001, a check is made to see if data distribution puts the emphasis on the expected degree of parallelism. If the emphasis is not found to be placed on the expected degree of parallelism, step 21002 is reached. If the emphasis is found to be placed on the expected degree of parallelism, step 21010 is reached.

In step 21002, a check is made to see if a user's designation exists. If the user's designation is found, step 21016 is reached. If the user's designation is not found, the current processing is terminated.

In step 21003, the number of necessary pages for storing the rows of the whole table is calculated.

Step 21004 decides the upper limit number of pages to be stored into disk units which are parallelly accessible within a predetermined time that is needed to scan the whole table.

In step 21005, the BES nodes, IOS nodes and

disk units which satisfy the preceding requirement are determined.

In step 21006, a check is made to see if there is a key range division. If there exists a key range division, step 21007 is reached. If there is no key range division, step 21009 is reached.

In step 21007, the divided key range is subdivided by an upper limit page count.

In step 21008, data is inserted into the key
10 range subdivisions. Thereafter, the current processing
is terminated.

In step 21009, data is inserted as partitioned by the upper limit page count. Then the current processing is terminated.

15 In step 21010, the optimum page access count m
is calculated on the basis of the estimated work load.

In step 21011, the expected degree of parallelism p is calculated, and the BES node, IOS node and disk units are determined accordingly.

20 In step 21012, a check is made to see if there exists a key range division. If there is a key range division, step 21013 is reached. If there is no key range division, step 21015 is reached.

25 In step 21013, the number of pages s to be stored in sub-key range units is calculated ($s = m/p$).

In step 21014, pages are divided in units of s pages into the sub-key ranges and data is inserted into each of the disk units. Then the current processing is terminated.

5 In step 21015, data is inserted as partitioned in units of s pages to each disk unit.

In step 21016, data is inserted to the disk units managed by the user-designated IOS node. The current processing is then terminated.

10 Fig. 22 is a flowchart of steps performed by the dynamic load control program 211. In step 21100, a check is made to see if there is a load unbalance (access concentrated or discrete). Specifically, the congested resources (processors (BES, IOS), disk units)
15 are detected from the database processing load executed in the unit time per node (i.e., the load is composed of the number of processing steps (for DB processing, I/O processing and communication processing), of the processor performance (converted to processing time),
20 and of the I/O operation count (converted to I/O time)). The DB processing is then translated into SQL sentences and the status of access to each resource is sorted in units of tables. If a load unbalance is detected, step 21101 is reached; if no load unbalance is detected, the
25 current processing is terminated.

5 necessary, step 21102 is reached; if no such addition or
deletion is necessary, the current processing is
terminated.

10 needed, step 21103 is reached; if no such addition is
needed, step 21112 is reached.

15 online processing is not in effect, step 21105 is
reached.

In step 21104, the key range of the table managed by the target BES nodes is closed.

In step 21105, new BES nodes are assigned.

20 In step 21106, lock information and directory
information are succeeded.

In step 21107, the DS 71 is ordered to update the dictionary information necessary for node assignment control.

25 In step 21108, a check is made to see if an IOS

node exists. If no IOS node is found, step 21109 is reached. If an IOS node is found to exist, step 21110 is reached. This step is inserted here so that the same dynamic load control program may address two kinds of
5 system configurations: one wherein the IOS exists and the other in which no IOS exists.

In step 21109, data is transferred from the target BES nodes to the new BES nodes.

In step 21110, a check is made to see if online
10 processing is in progress. If online processing is found to be in progress, step 21111 is reached. If online processing is not in effect, the current processing is terminated.

In step 21111, the closing of the key range of
15 the table managed by the target BES nodes is released. Then the current processing is terminated.

In step 21112, a check is made to see if online processing is in progress. If online processing is found to be in progress, step 21113 is reached. If
20 online processing is not in effect, step 21114 is reached.

In step 21113, the key range of the table managed by the target BES nodes is closed.

In step 21114, the BES nodes to be degenerated
25 are determined.

In step 21115, the lock information and directory information are succeeded.

In step 21116, the DS 71 is ordered to update the dictionary information necessary for node assignment control.

In step 21117, a check is made to see if an IOS node exists. If no IOS node is found, step 21118 is reached. If an IOS node is found to exist, step 21119 is reached.

10 In step 21118, data is purged out of the BES
nodes to be degenerated.

In step 21119, a check is made to see if online processing is in progress. If online processing is found to be in progress, step 21120 is reached. If
15 online processing is not in effect, the current processing is terminated.

In step 21120, the closing of the key range of the table managed by the target BES nodes is released. Then the current processing is terminated.

20 Fig. 23 is a conceptual view of dynamic data
load control using the key range division scheme. For
this setup, it is assumed that the partition count is 4
and that the column values v1 - v6 of the database take
the occurrence frequencies of Fig. 11.

25 At initial data load time, only one BES node

731 is needed. The upper limit number of pages to be stored is made to correspond with each of the partitions 810 - 840 constituted by the disk units. In this setup, the column values v1 - v2 are stored in the disk units
5 of the partition 810; the column values v2 - v3 are stored in the disk units of the partitions 820 - 830; the column values v3 - v5 are stored in the disk units of the partition 840; and the column values v5 - v6 are stored in other disk units. At initial data load time,
10 the directory information is generated for each disk unit so as to manage the pages stored therein.

Where the BES nodes 732 - 734 are used selectively in accordance with the load, the directory information regarding each disk unit corresponding to
15 the configured BES nodes is utilized for access to the database.

The processes described above may be implemented in other variations of the invention outlined below. In one variation, information about BES
20 positions is not included in row identifiers so as to facilitate the transfer of rows between nodes. A BES node determines the physical position of a given row by combining the row identifier with the directory information for defining the divided positions of the
25 table. Row transfer is accomplished by updating the

directory information. There may be provided a structure capable of dealing with node reconfiguration or row transfer, the structure allowing the processing to be partitioned through succession of the directory
5 information and lock information even as BES nodes are added dynamically.

If it is desired to manage the database in a replica setup, the storage area needs to be doubled. This means that the disk access load is approximately
10 doubled whether or not primary and backup copies are managed by the same IOS and BES nodes. Therefore the number of volumes for each partition managed with the exist partition count need only be halved.

In case of a disk unit, IOS or BES failure, the
15 failed unit or node is disconnected from the online processing. The failed unit or node is repaired and then reconnected to the online processing. The way in which the closing of the key range is managed varies with the node. Specifically, in case of a disk unit
20 failure, the key range stored in that disk unit is closed. If a backup copy exists (the backup copy needs to be acquired under management of the same IOS node (mirror disk unit) or of another IOS node (data replica)), the processing is reassigned. In case of an
25 IOS node failure, the key range stored in that IOS node

is closed. If a backup copy exists (the backup copy needs to be acquired under management of another IOS node (data replica)), the processing is reassigned. In case of a BES node failure, the key range managed by
5 that BES node is closed. If an IOS node exists, new BES nodes are assigned, lock information is succeeded, the dictionary information necessary for node assignment control is updated, and the processing is allowed to continue.

10 This invention is not limited to systems wherein rules based on statistical information and cost evaluations are used in combination. The invention may also be applied to database management systems performing optimization processing by use of cost
15 evaluations alone, the rules alone, or the combination of cost evaluations and the rules, as long as processing procedures offering appropriate database reference characteristic information can be acquired thereby.

 This invention may be practiced via a software
20 system for a tightly or loosely coupled composite processor system in a mainframe computer. It is also possible to practice the invention via a tightly or loosely coupled composite processor system using a dedicated processor for each of the component processing
25 programs. The invention may also be applied to a single

processor system if the system assigns parallel processes to each processing procedure.

Furthermore, the invention may be applied to configurations wherein a plurality of processors share
5 each of a plurality of disk units.

The database divisional management method of the invention renders the system configuration suitable for the load in question. This provides the desired degree of parallelism and permits high-speed query processing.

The parallel database system according to the invention is thus a scalable parallel database system capable of altering the system configuration constantly in keeping with any fluctuation in load.

15 As many apparently different embodiments of this invention may be made without departing from the spirit and scope thereof, it is to be understood that the invention is not limited to the specific embodiments thereof except as defined in the appended claims.

WHAT IS CLAIMED IS:

1. A database divisional management method for use with a parallel database system having a storage medium, storage and management means for storing and managing a database in said storage medium, and a plurality of access means for accessing said database in response to query inputs, said database divisional management method comprising the steps of:

```

                                generating a procedure for processing said
10  query inputs;

```

dividing said database into a plurality of partitions in accordance with the load pattern provided for executing the generated processing procedure; and

determining the number of access means for
15 simultaneously accessing the partitions of said
database.

2. A database divisional management method according to claim 1, wherein said storage and management means determines the physical addresses corresponding to logical addresses at which said plurality of access means access said partitions of said database.

3. A database divisional management method according to claim 1, wherein said load pattern is
25 determined by the access efficiency of each of said

access means and by the amount of information stored in said partitions of said database accessed by said access means.

4. A database divisional management method for
5 use with a parallel database system comprising an FES
node for generating a processing procedure in response
to query input information, BES nodes for accessing a
database on the basis of said processing procedure
generated by said FES node, an IOS node having a storage
10 medium and capable of storing and managing said database
in said storage medium, and a network for connecting the
FES, BES and IOS nodes, said database divisional
management method comprising the steps of:

calculating the load pattern by which to
15 perform database processing using said processing
procedure; and

determining the number of processors assigned
to said FES node, the number of processors assigned to
said BES nodes and the number of processors assigned to
20 said IOS node in accordance with said load pattern.

5. A database divisional management method
according to claim 4, further comprising the step of
determining the number of storage medium units of said
IOS node and the number of partitions in each of said
25 storage medium units in accordance with said load

6. A database divisional management method according to claim 5, wherein said storage medium units are disk units and wherein the number of disk units of said IOS node and the number of partitions of said disk units are determined in accordance with said load pattern.

8. A database divisional management method for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes having a storage medium in which to store a database and capable of accessing said database on the basis of said processing procedure generated by said FES node, and a network for connecting the FES and BES nodes, said database divisional management method comprising the steps of:

determining the number of processors assigned to said FES node, the number of processors assigned to

said BES nodes, the number of storage medium units of said BES nodes, and the number of partitions of said storage medium in accordance with said load pattern.

9. A database divisional management method for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes for accessing a database on the basis of said processing procedure generated by said FES node, an IOS node having a storage medium and capable of storing and managing said database in said storage medium, and a network for connecting the FES, BES and IOS nodes, said database divisional management method comprising the steps of:

determining the upper limit number of pages which are accessible in parallel and which require a constant time each when said database is scanned for access thereto; and

determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes, and the number of processors assigned to said IOS node in accordance with said upper limit number of pages.

10. A database divisional management method according to claim 9, further comprising the step of determining the number of storage medium units of said

IOS nodes and the number of partitions in each of said storage medium units.

11. A database divisional management method for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes having a storage medium in which to store and manage a database and capable of accessing said database on the basis of said processing procedure generated by said FES node, and a network for connecting the FES and BES nodes, said database divisional management method comprising the steps of:

determining the upper limit number of pages which are accessible in parallel and which require a constant time each when said database is scanned for access thereto; and

determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes, the number of storage medium units of said BES nodes, and the number of partitions of said storage medium units in accordance with said upper limit number of pages.

12. A database divisional management method for use with a parallel database system comprising an FES node for generating a processing procedure in

response to query input information, BES nodes for
accessing a database on the basis of said processing
procedure generated by said FES node, an IOS node having
a storage medium and capable of storing and managing
5 said database in said storage medium, and a network for
connecting the FES, BES and IOS nodes, said database
divisional management method comprising the steps of:

calculating the expected degree of parallelism
p according to the load pattern based on said processing
10 procedure; and

determining the number of processors assigned
to said FES node, the number of processors assigned to
said BES nodes, and the number of processors assigned to
said IOS node in accordance with said expected degree of
15 parallelism p.

13. A database divisional management method
for use with a parallel database system comprising an
FES node for generating a processing procedure in
response to query input information, BES nodes having a
20 storage medium in which to store and manage a database
and capable of accessing said database on the basis of
said processing procedure generated by said FES node,
and a network for connecting the FES and BES nodes, said
database divisional management method comprising the
25 steps of:

calculating the expected degree of parallelism
p according to the load pattern based on said processing
procedure; and

determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes, the number of storage medium units of said BES nodes, and the number of partitions of said storage medium units in accordance with said expected degree of parallelism p.

10 14. A database divisional management method
according to claim 4, further comprising the steps of:
calculating the optimum number of accessible
pages m;

calculating the number of pages s ($=m/p$,
15 wherein p denotes an expected degree of parallelism) in
units of sub-key ranges if key range partitions exist;
and

having sub-key range partitions in units of s
pages for inserting data into a disk apparatus.

20 15. A database divisional management method
according to claim 8, further comprising the steps of:
calculating the optimum number of accessible
pages m;

calculating the number of pages s ($=m/p$,
25 wherein p denotes an expected degree of parallelism) in

units of sub-key ranges if key range partitions exist;
and

having sub-key range partitions in units of s
pages for inserting data into a disk apparatus.

5 16. A database divisional management method
according to claim 9, further comprising the steps of:
calculating the optimum number of accessible
pages m;

calculating the number of pages $s (=m/p$,
10 wherein p denotes an expected degree of parallelism) in
units of sub-key ranges if key range partitions exist;
and

having sub-key range partitions in units of s
pages for inserting data into a disk apparatus.

15 17. A database divisional management method
according to claim 11, further comprising the steps of:
calculating the optimum number of accessible
pages m;

calculating the number of pages s ($=m/p$,
20 wherein p denotes an expected degree of parallelism) in
units of sub-key ranges if key range partitions exist;
and

having sub-key range partitions in units of s
pages for inserting data into a disk apparatus.

25 18. A database divisional management method

calculating the optimum number of accessible
pages m;

5

10

according to claim 13, further comprising the steps of:

pages m;

15

20

25

accessing a database on the basis of said processing
procedure generated by said FES node, an IOS node having
a storage medium and capable of storing and managing
said database in said storage medium, and a network for
5 connecting the FES, BES and IOS nodes, said database
divisional management method comprising the steps of:

detecting a load unbalance on the basis of at
least one of the load information items consisting of
the number of accessed pages, the number of hit rows and
10 the number of communications acquired during execution
of said processing procedure; and

changing the number of processors assigned to
said FES node, the number of processors assigned to said
BES nodes, the number of processors assigned to said IOS
15 node, and the number of storage medium units of said IOS
node so as to eliminate said load unbalance.

21. A database divisional management method
for use with a parallel database system comprising an
FES node for generating a processing procedure in
20 response to query input information, BES nodes having a
storage medium in which to store and manage a database
and capable of accessing said database on the basis of
said processing procedure generated by said FES node,
and a network for connecting the FES and BES nodes, said
25 database divisional management method comprising the

for node assignment control; and

releasing the closing of said key range
thereafter if said online processing is still in
progress.

5 23. A database divisional management method
according to claim 21, further comprising the steps of:

 closing, when online processing is in progress,
the key range of a database table if either the number
of processors assigned to said BES nodes or the number
10 of storage medium units is to be increased, said
database table being the object to be managed by either
the processors or the storage medium units to be added;

 assigning either the processors or the storage
medium units anew;

15 succeeding lock information and directory
information;

 updating the dictionary information necessary
for node assignment control;

 moving data from the existing group of storage
20 medium units to the newly added storage medium units;
and

 releasing the closing of said key range
thereafter if said online processing is still in
progress.

25 24. A database divisional management method

according to claim 20, further comprising the steps of:

5 closing, when online processing is in progress,
the key range of a database table if at least one of the
three numbers consisting of the number of processors
assigned to said BES nodes, the number of processors
assigned to said IOS node and the number of storage
medium units is to be decreased, said database table
being managed by either the processors or the storage
medium units to be removed;

10 determining either the processors or the
storage medium units to be removed;

succeeding lock information and directory
information;

15 updating the dictionary information necessary
for node assignment control; and

releasing the closing of said key range
thereafter if said online processing is still in
progress.

25. A database divisional management method
20 according to claim 22, further comprising the steps of:

25 closing, when online processing is in progress,
the key range of a database table if at least one of the
three numbers consisting of the number of processors
assigned to said BES nodes, the number of processors
assigned to said IOS node and the number of storage

updating the dictionary information necessary


```
for node assignment control;
```

moving data from the storage medium units to be removed to the storage medium units succeeding those units to be removed; and

5 releasing the closing of said key range
thereafter if said online processing is still in
progress.

27. A database divisional management method according to claim 23, further comprising the steps of

10 closing, when online processing is in progress,
the key range of a database table if at least either the
number of processors assigned to said BES nodes or the
number of storage medium units is to be decreased, said
database table being managed by either the processors or
15 the storage medium units to be removed;

determining either the processors or the storage medium units to be removed;

```

        succeeding lock information and directory
information;

```

```
20      updating the dictionary information necessary
      for node assignment control;
```

moving data from the storage medium units to be removed to the storage medium units succeeding those units to be removed; and

25 releasing the closing of said key range

said processing procedure; and

determination means for determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes and the number of processors assigned to said IOS node in accordance with said load pattern.

35. A database divisional management system for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes having a storage medium in which to store a database and capable of accessing said database on the basis of said processing procedure generated by said FES node, and a network for connecting the FES and BES nodes, said database divisional management system comprising:

calculation means for calculating the load pattern by which to perform database processing using said processing procedure; and

determination means for determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes, the number of storage medium units of said BES nodes, and the number of partitions of said storage medium in accordance with said load pattern.

36. A database divisional management system

for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes for accessing a database on the basis of said processing
5 procedure generated by said FES node, an IOS node having a storage medium and capable of storing and managing said database in said storage medium, and a network for connecting the FES, BES and IOS nodes, said database divisional management system comprising:

10 determination means for determining the upper limit number of pages which are accessible in parallel and which require a constant time each when said database is scanned for access thereto; and

determination means for determining the number
15 of processors assigned to said FES node, the number of processors assigned to said BES nodes, and the number of processors assigned to said IOS node in accordance with said upper limit number of pages.

37. A database divisional management system
20 for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes having a storage medium in which to store and manage a database and capable of accessing said database on the basis of
25 said processing procedure generated by said FES node,

determination means for determining the upper limit number of pages which are accessible in parallel

5 and which require a constant time each when said
database is scanned for access thereto; and

determination means for determining the number of processors assigned to said FES node, the number of processors assigned to said BES nodes, the number of storage medium units of said BES nodes, and the number of partitions of said storage medium units in accordance with said upper limit number of pages.

38. A database divisional management system for use with a parallel database system comprising an FES node for generating a processing procedure in response to query input information, BES nodes for accessing a database on the basis of said processing procedure generated by said FES node, an IOS node having a storage medium and capable of storing and managing said database in said storage medium, and a network for connecting the FES, BES and IOS nodes, said database divisional management system comprising:

calculation means for calculating the expected
degree of parallelism p according to the load pattern
25 based on said processing procedure; and

ABSTRACT OF THE DISCLOSURE

A method and a system of database divisional management for use with a parallel database system comprising an FES (front end server), BES's (back end servers), an IOS (I/O server) and disk units. The numbers of processors assigned to the FES, BES's and IOS, the number of disk units, and the number of partitions of the disk units are determined in accordance with the load pattern in question. Illustratively, there may be established a configuration of one FES, four BES's, one IOS and eight disk units. The number of BES's is varied from one to four depending on the fluctuation in load, so that a scalable system configuration is implemented. When the number of BES's is increased or decreased, only the management information thereabout is transferred between nodes and not the data, whereby the desired degree of parallelism is obtained for high-speed query processing.

000000-000000

FIG. 1

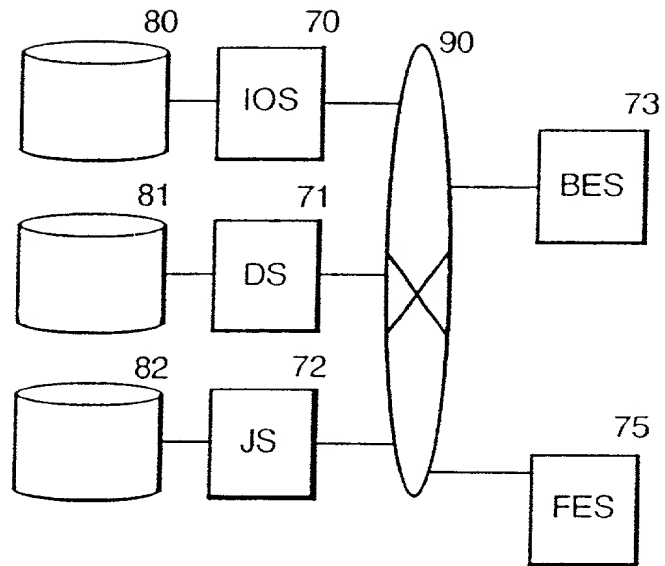


FIG. 2

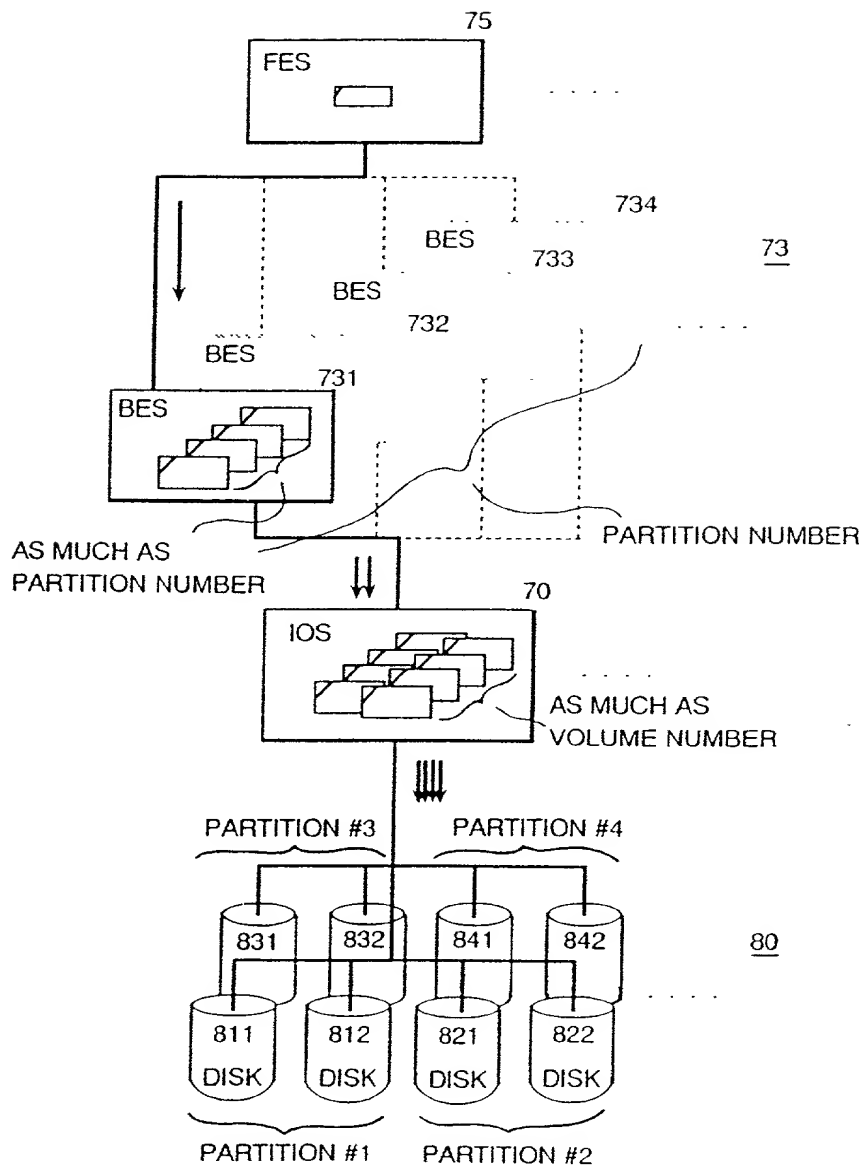
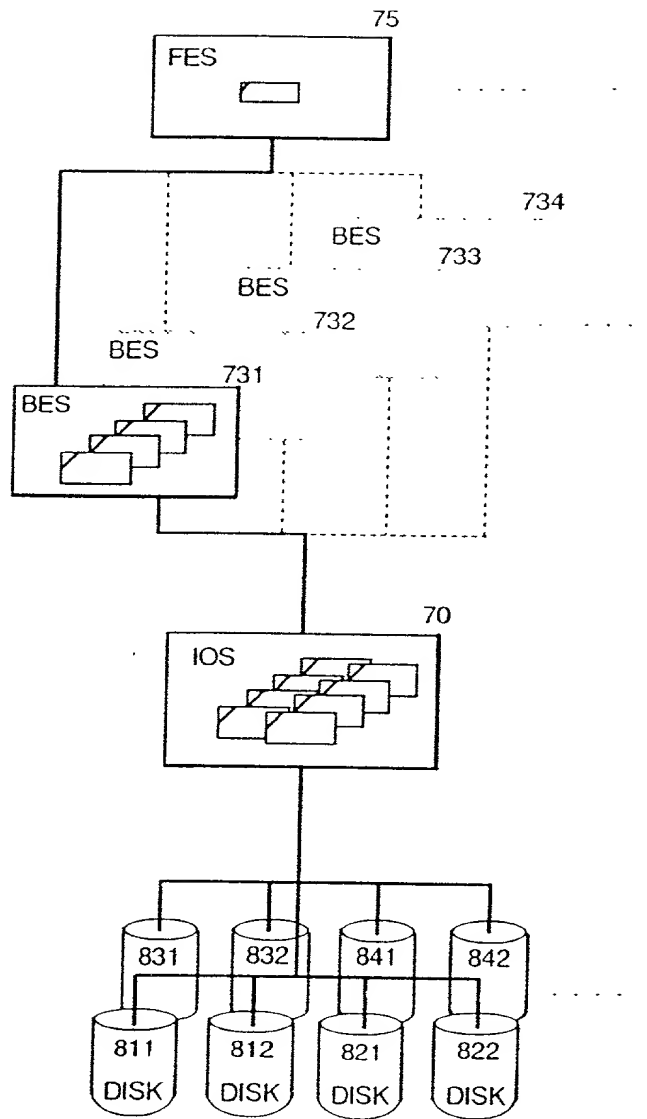


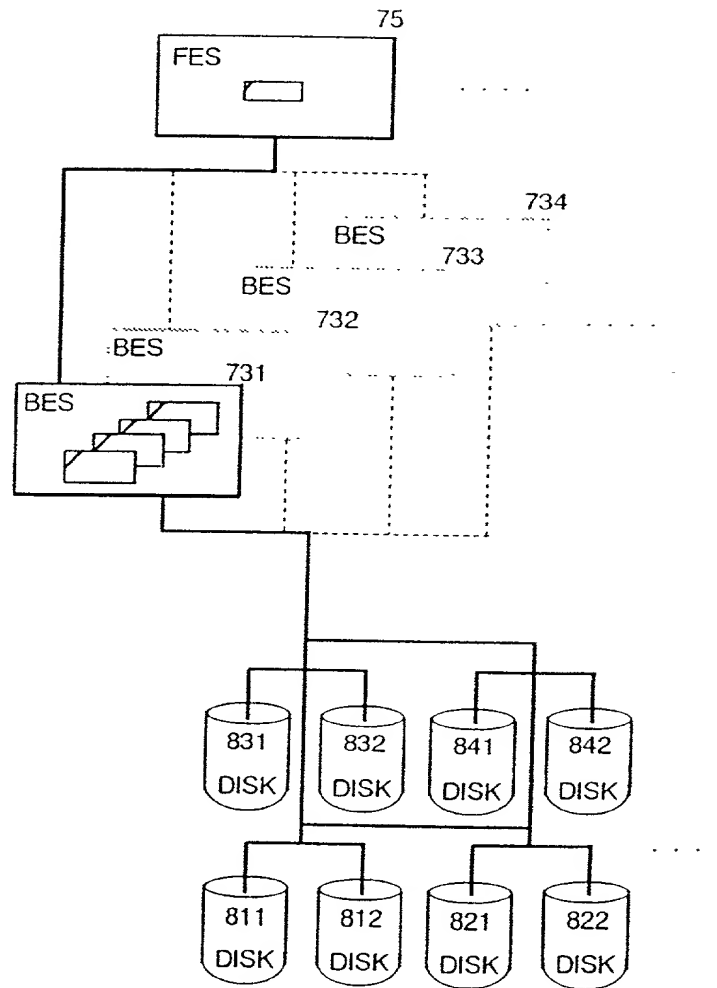
FIG. 3



OPTIMAL NODE DISTRIBUTION :

FES . BES : IOS . DISK
 = 1 : 4 . 1 . 8

FIG. 4



OPTIMAL NODE DISTRIBUTION :

FES . BES . DISK
= 1 . 4 : 8

FIG. 5

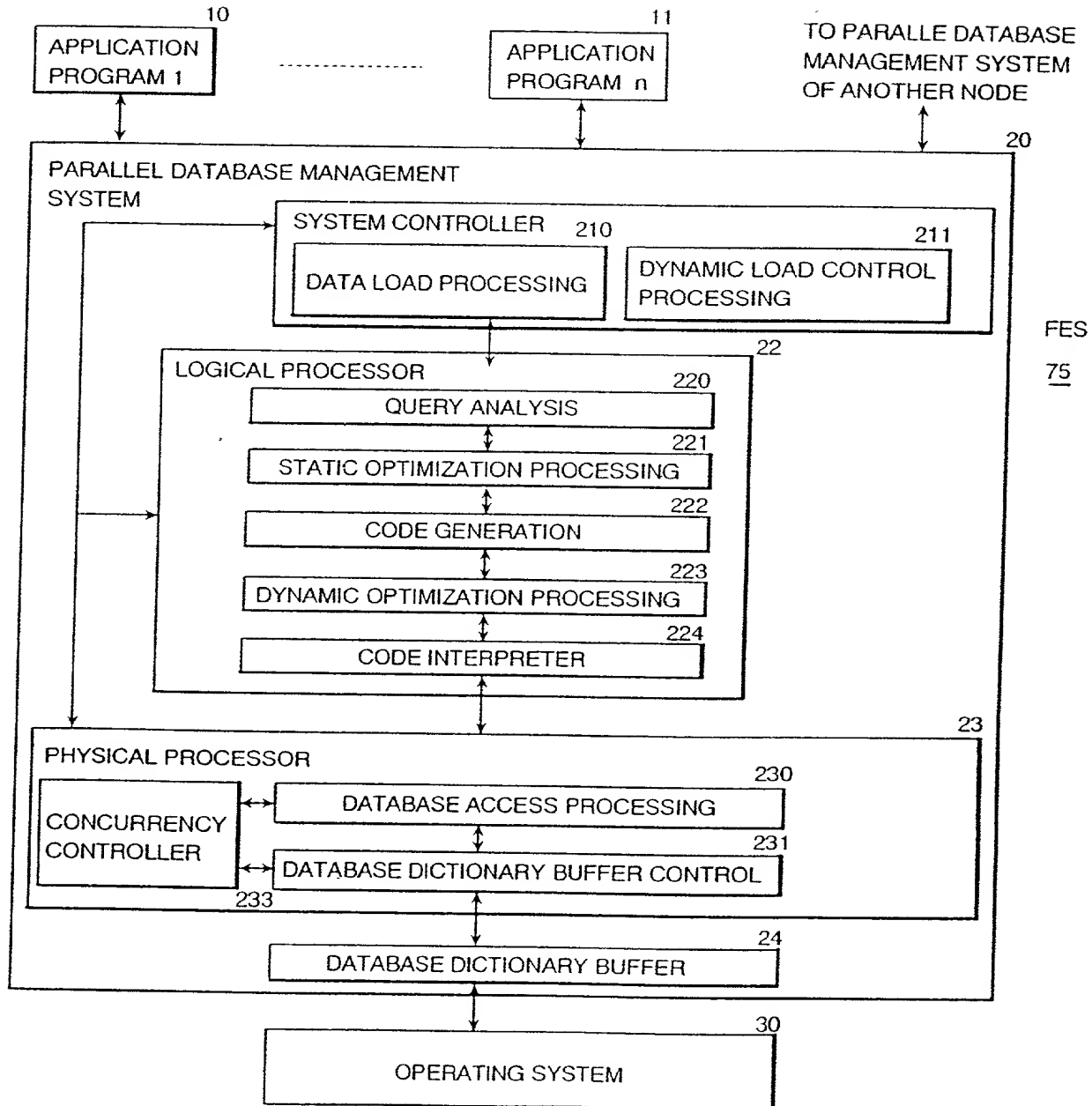


FIG. 6

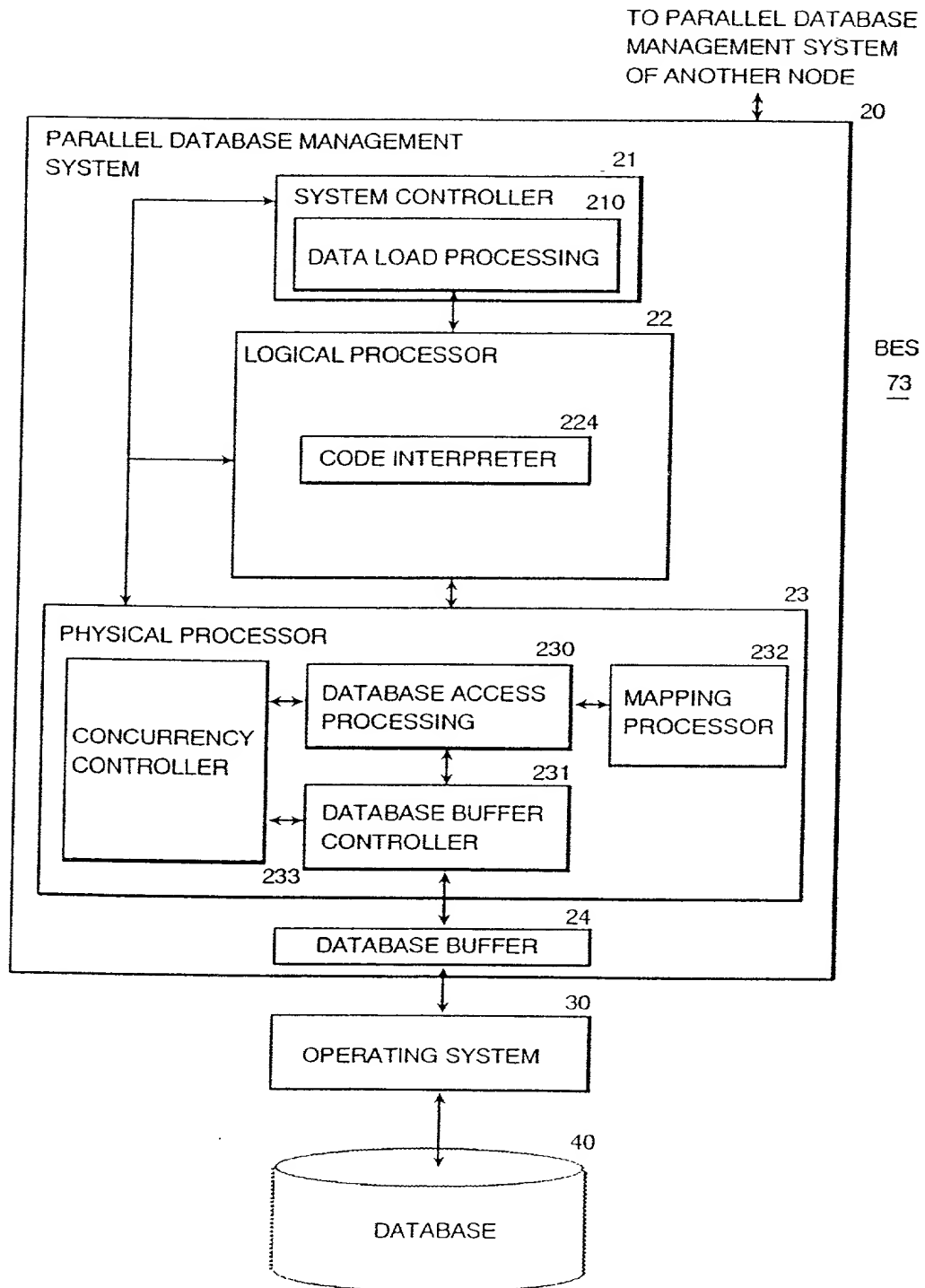


FIG. 7

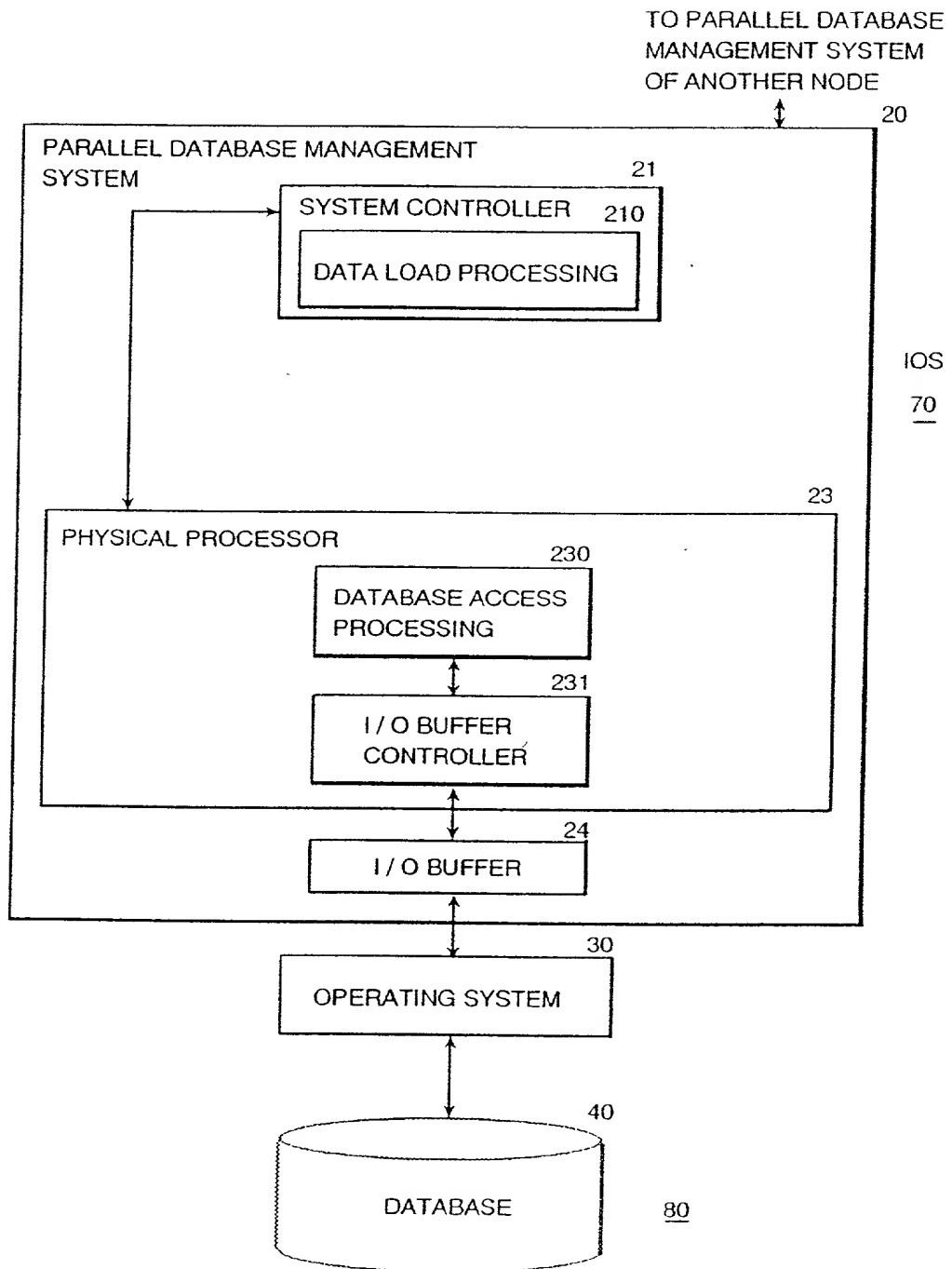
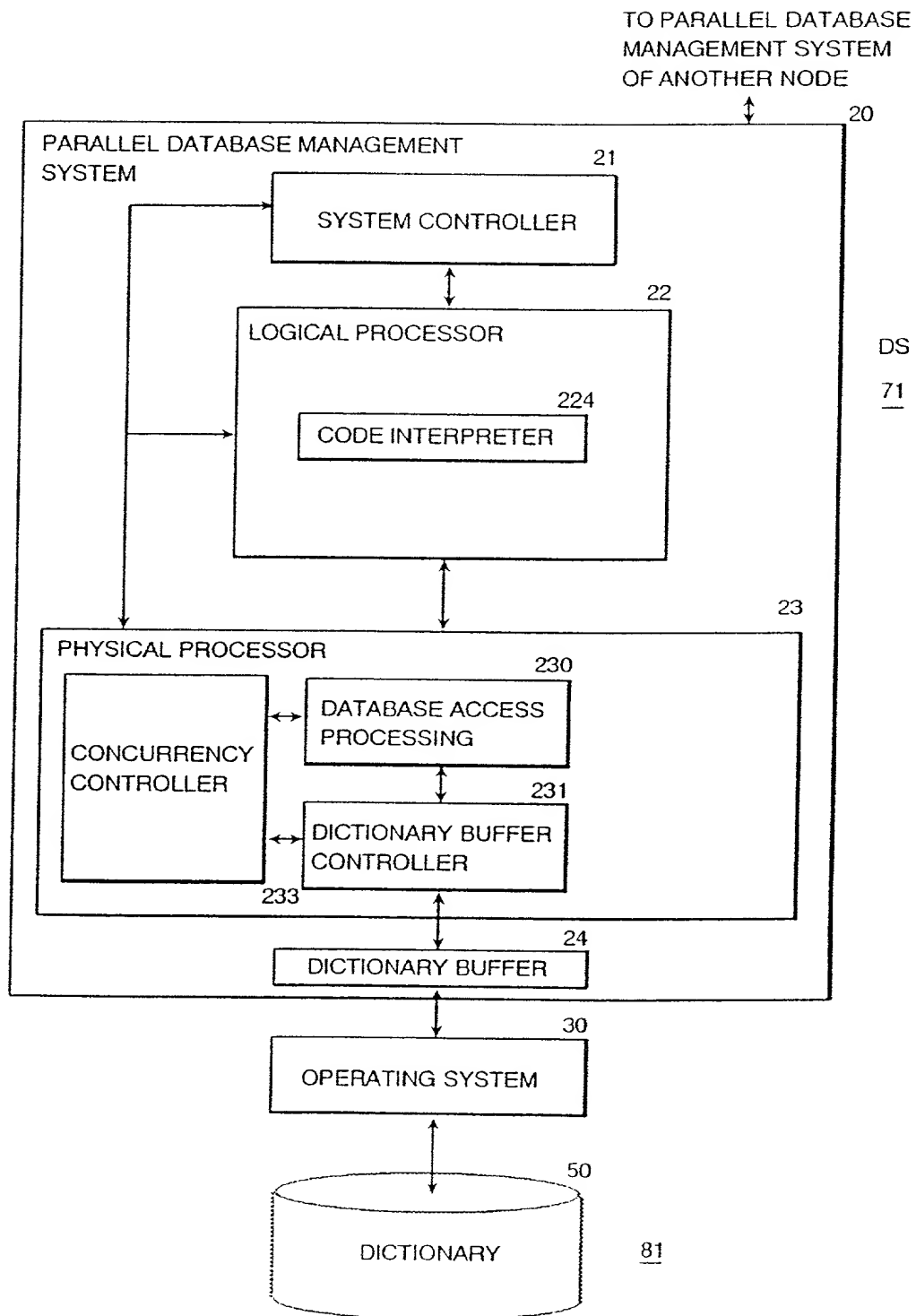


FIG. 8



DS
71

FIG. 9

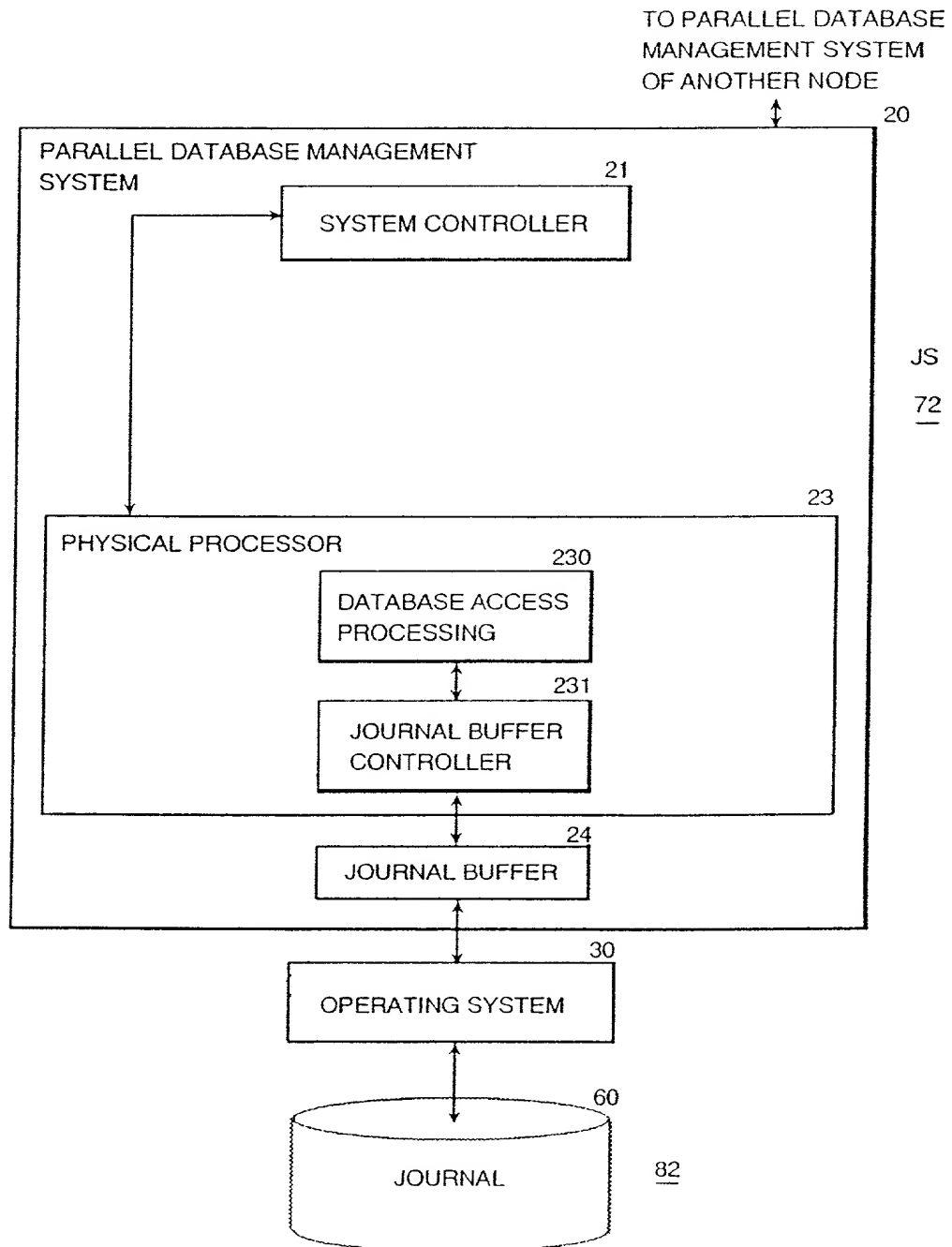


FIG. 11

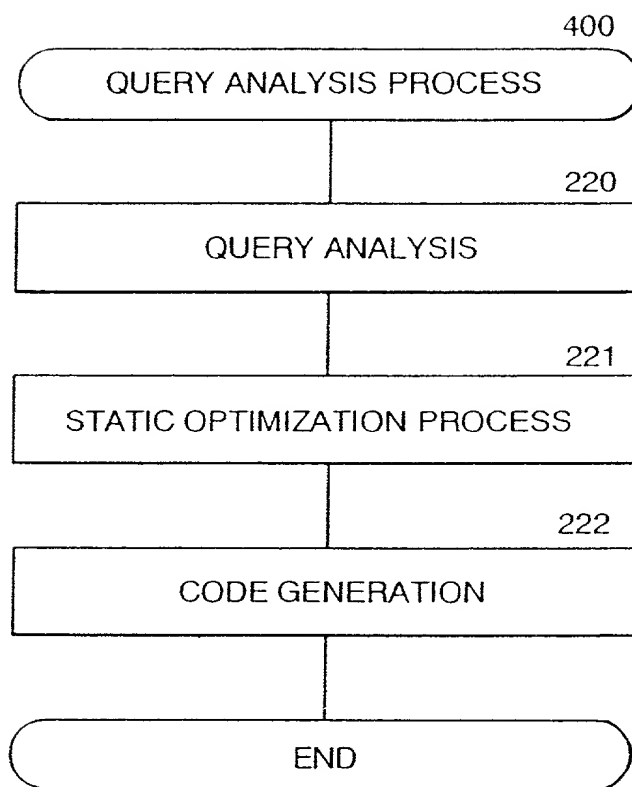


FIG. 12

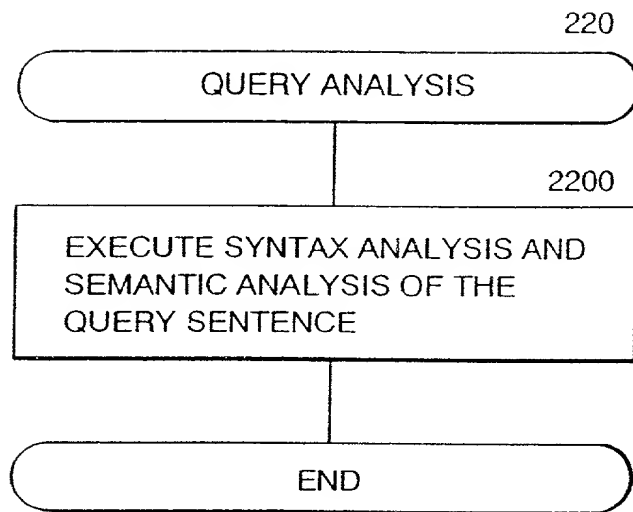


FIG. 13

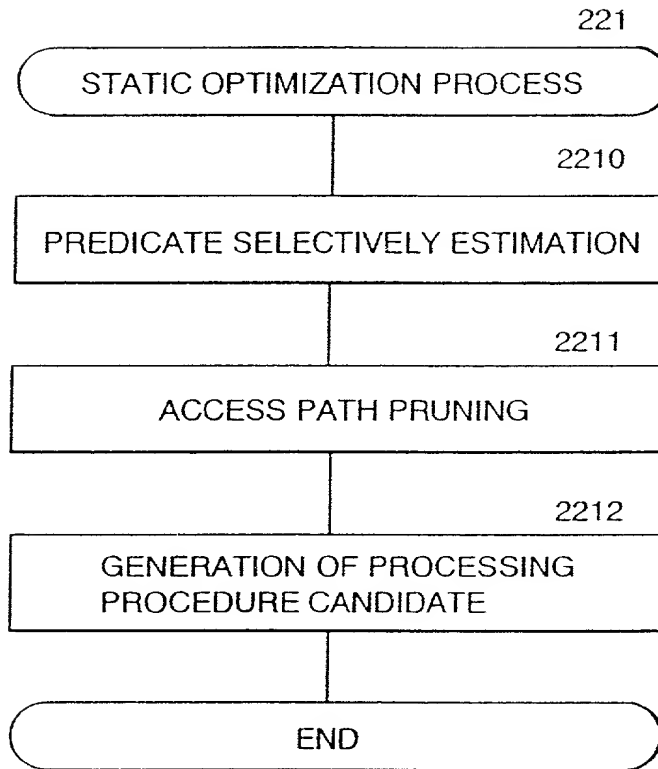


FIG. 14

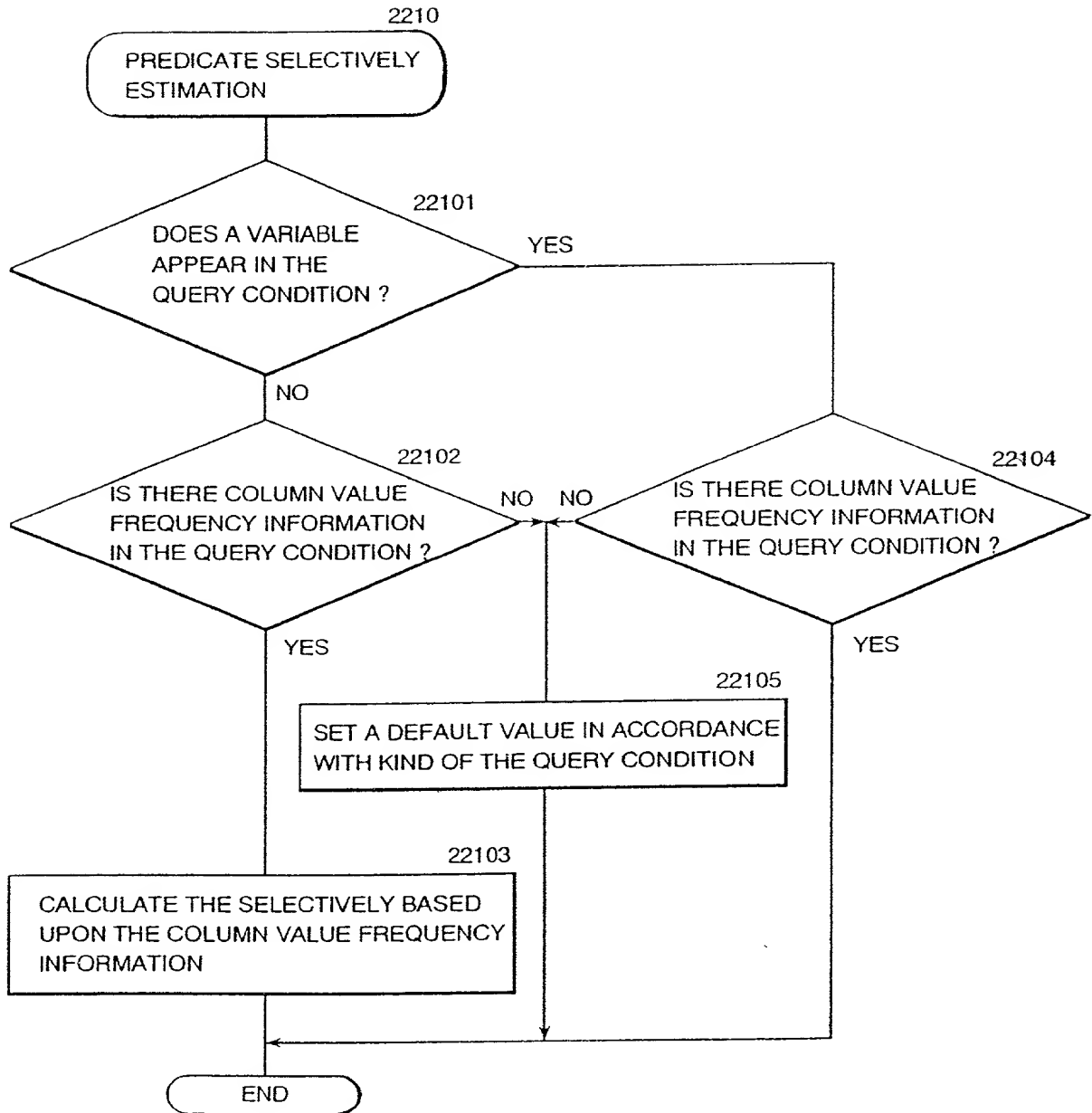


FIG. 15

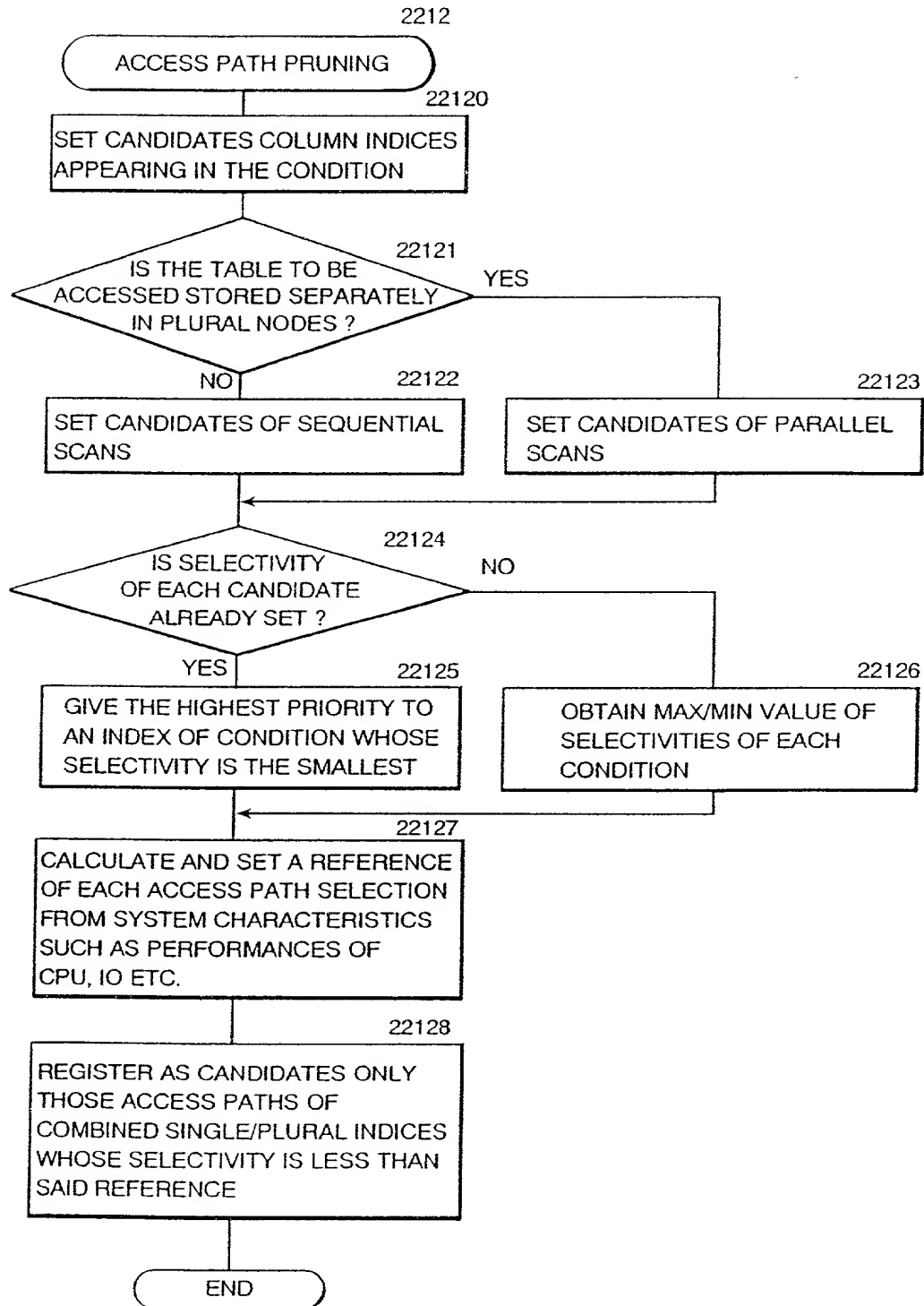


FIG. 16

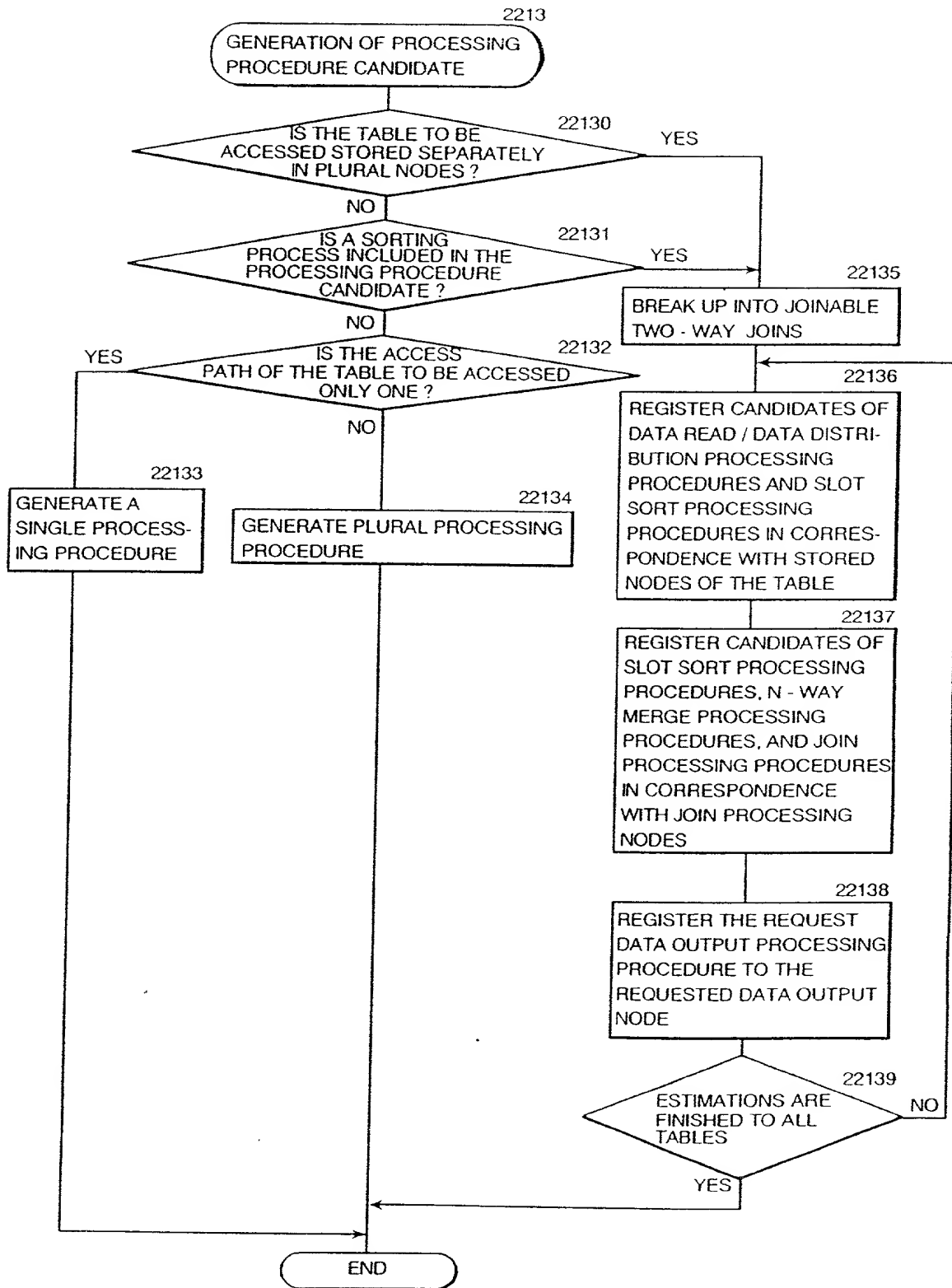


FIG. 17

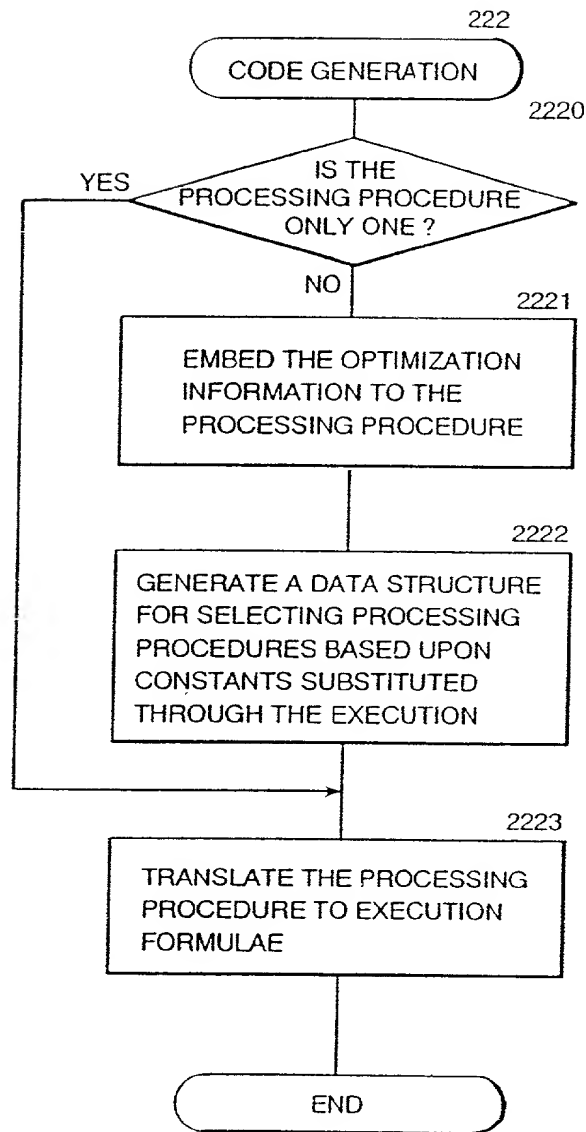


FIG. 18

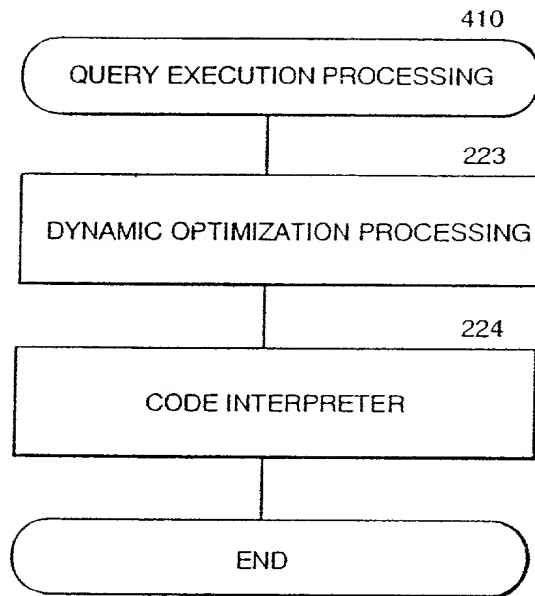


FIG. 19

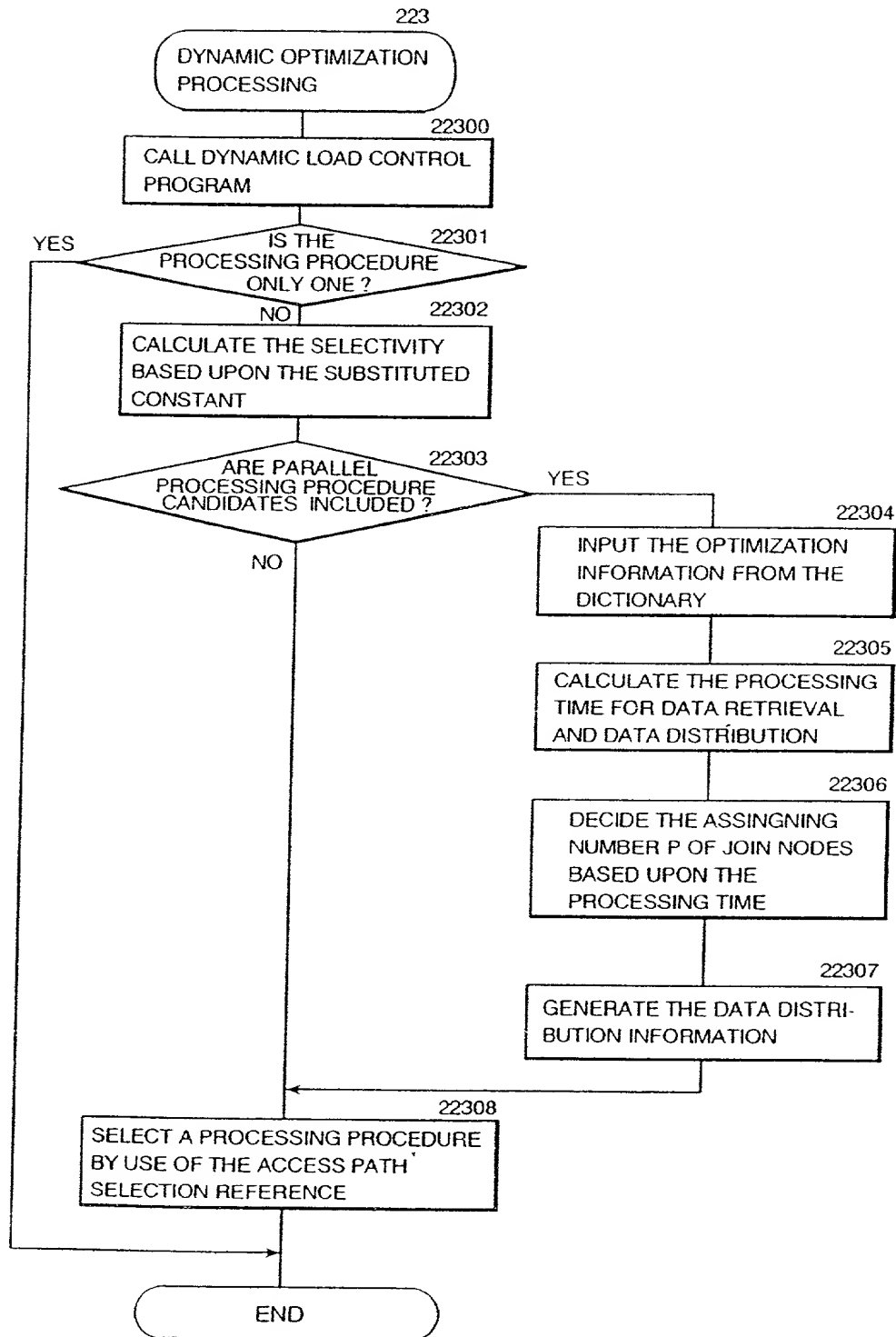


FIG. 20

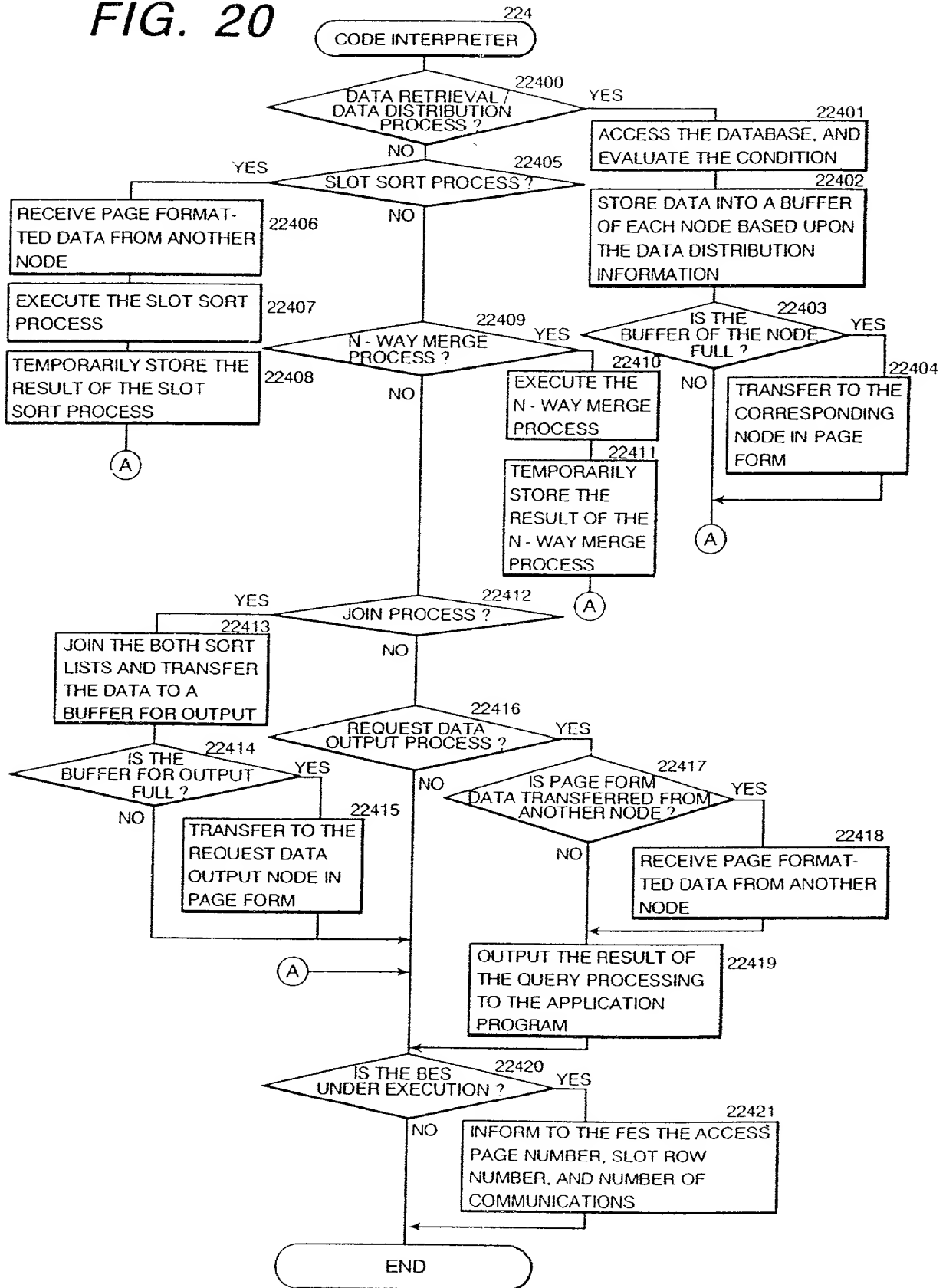


FIG. 21

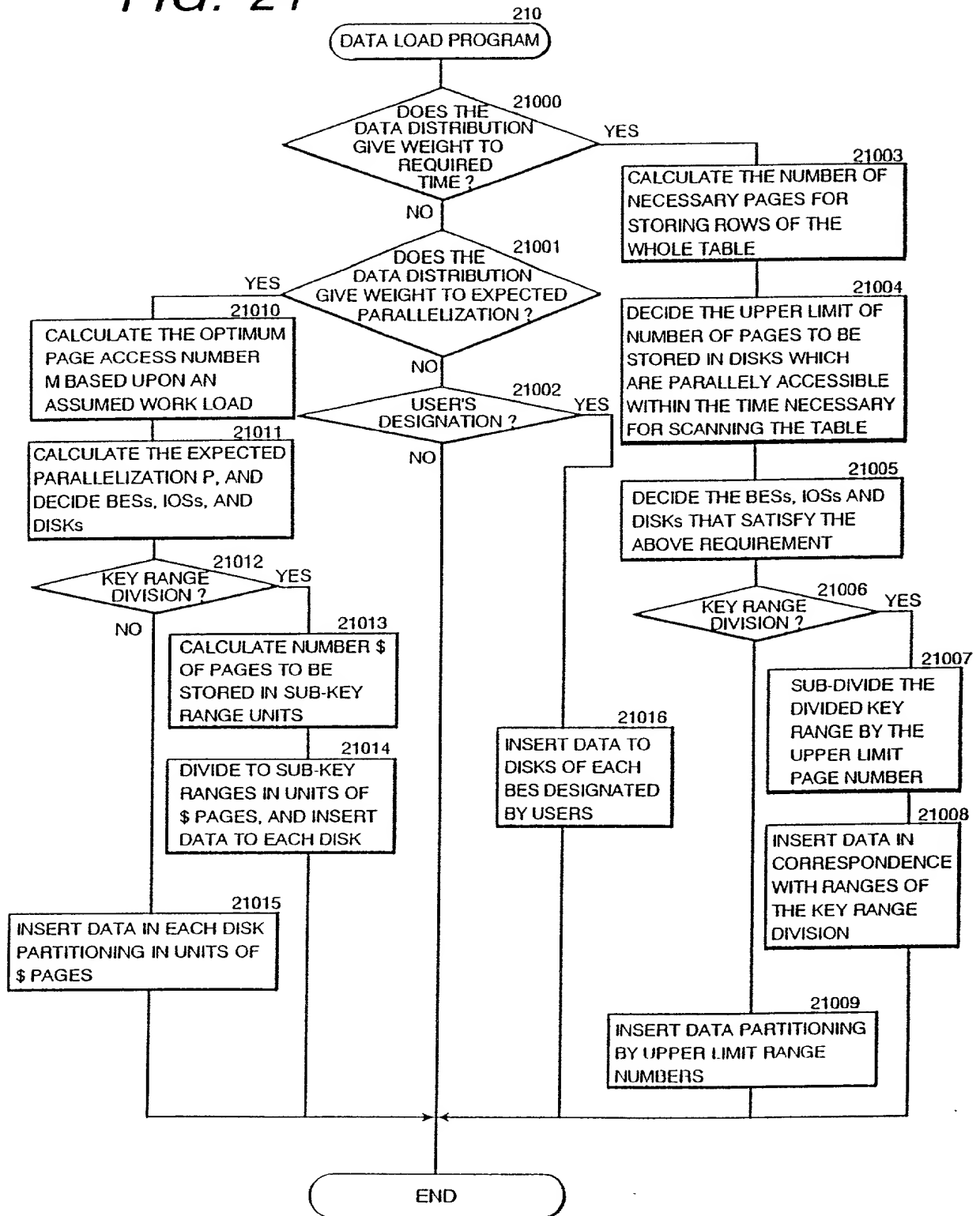


FIG. 22

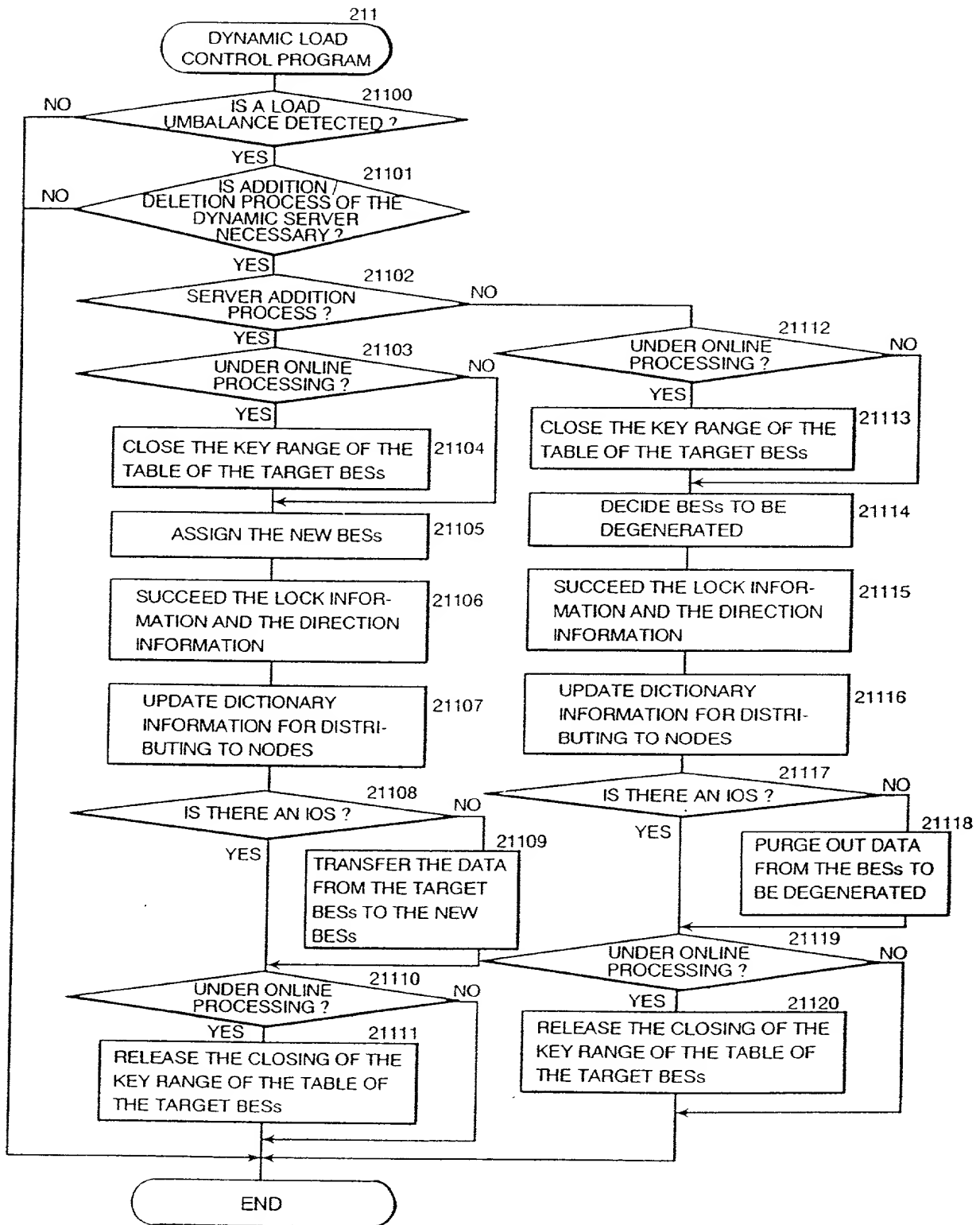
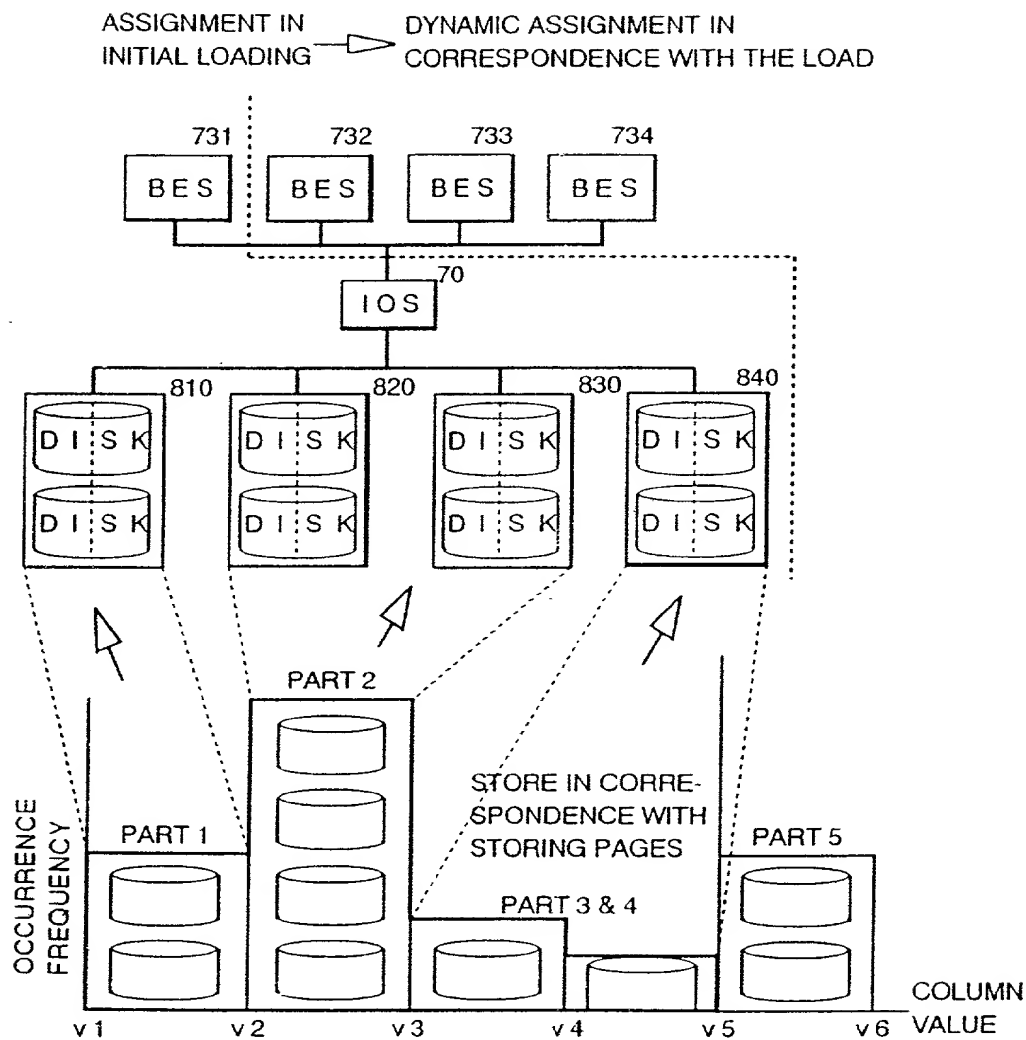


FIG. 23



COMBINED DECLARATION AND POWER OF ATTORNEY

(宣誓書及び委任状)

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name, I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

METHOD AND SYSTEM OF DATABASE DIVISIONAL MANAGEMENT FOR
PARALLEL DATABASE SYSTEM

the specification of which: (check one) ☒ is attached hereto.

☐ was filed on _____
as Application Serial No. _____
and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)			Priority Claimed	
<u>5-286549</u> (Number)	<u>Japan</u> (Country)	<u>16/Nov./1993</u> (Day/Month/Year Filed)	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No

I hereby claim the benefit under Title 35, United States Code, §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status) (patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status) (patented, pending, abandoned)

Antonelli, Terry, Stout & Kraus
Suite 1800
1300 North Seventeenth Street
Arlington, Virginia 22209
Telephone: (703) 312-6600

目 次

発明者フルネームサイン

氏名タイプ欄

Residence Sagamihara-shi, Japan Citizenship Japan

Post Office Address #A204, Hitachi Kamitsuruma-shitaku, 17-12, Yutaka-cho,
Sagamihara-shi, Kanagawa-ken 228, Japan

Residence Yokohama-shi, Japan Citizenship Japan

Post Office Address 7-31-14, Hino, Kounan-ku, Yokohama-shi, Kanagawa-ken 233, Japan

Date October 26, 1994 Inventor Shunichi Torii Shunichi Torii

Residence Musashino-shi, Japan Citizenship Japan

Post Office Address #8-125, 3-7, Naka-machi, Musashino-shi, Tokyo 180, Japan

Date _____ Inventor _____

Residence _____ Citizenship _____

Post Office Address _____

Date _____ Inventor _____

Residence _____ Citizenship _____

Post Office Address _____

Date _____ Inventor _____

Residence _____ Citizenship _____

Post Office Address _____